DOCUMENT RESUME

ED 429 093 TM 029 630

AUTHOR	Glass, Gene V., Ed.
TITLE	Education Policy Analysis Archives, 1998. Volume 6, 1998.
ISSN	ISSN-1068-2341
PUB DATE	1998-00-00
NOTE	755p.; "Education Policy Analysis Archives" is an electronic-only journal covered on an article-by-article basis in "Current Index to Journals in Education" (CIJE). For the 21 articles in this volume, see TM 521 716-736 (in CIJE).
AVAILABLE FROM	Web site: http://olam.ed.asu.edu/epaa
PUB TYPE	Collected Works - Serials (022)
JOURNAL CIT	Education Policy Analysis Archives; v6 n1-21 1998
EDRS PRICE	MF04/PC31 Plus Postage.
DESCRIPTORS	*Academic Achievement; *Computer Uses in Education; Distance Education; *Educational Assessment; Educational Change; Educational Policy; Electronic Journals; Elementary Secondary Education; Foreign Countries; Higher Education; Internet; *Policy Formation; Scholarly Journals; *Standards
IDENTIFIERS	China; Greece; Latin America; Taiwan; Turkey

ABSTRACT

This document consists of printouts of downloaded copies of the 21 papers published in the electronic journal "Education Policy Analysis Archives" during 1996. The papers are: (1) "The Political Legacy of School Accountability Systems" (Sherman Dorn); (2) "Review of Stephen Arons's 'Short Route to Chaos'" (Charles L. Glenn); (3) "Planting Land Mines in Common Ground: A Review of Charles Glenn's Review of 'Short Route to Chaos'" (Stephen Arons); (4) "Comparative Issues of Selection in Europe: The Case of Greece" (Dionysias Gouvias); (5) "A Remarkable Move of Restructuring: Chinese Higher Education" (Fang Zhao); (6) "School Improvement Policy: Have Administrative Functions of Principals Changed in Schools Where Site Based Management Is Practiced?" (C. Kenneth Tanner and Cheryl D. Stone); (7) "Educational Research in Latin America: Review and Perspectives" (Abdeljalil Akkari and Soledad Perez); (8) "'The Art of Punishing': The Research Assessment Exercise and the Ritualisation of Power in Higher Education" (Lee-Anne Broadhead and Sean Howard); (9) "SOCRATES Invades Central Europe" (Joseph Slowinski); (10) "Educational Standards and the Problem of Error" (Noel Wilson); (11) "Public Policy on Distance Learning in Higher Education: California State and Western Governors Association Initiatives" (Gary A. Berg); (12) "Counseling in Turkey: Current Status and Future Challenges" (Suleyman Dogan); (13) "Consequences of Assessment: What Is the Evidence?" (William A. Mehrens); (14) "Some Comments on Assessment in U.S. Education" (Robert Stake); (15) "A Note on the Empirical Futility of Labor-Intensive Scoring Permutations for Assessing Scholarly Productivity: Implications for Research, Promotion/Tenure, and Mentoring" (Christine Hanish, John J. Horan, Bethanne Keen, and Ginger Clark); (16) "Criticizing the Schools: Then and Now" (Benjamin Levin); (17) "Performance Indicators: Information in Search of a Valid and Reliable Use" (E. Raymond Hackett and Sarah D. Carrigan); (18) "Transformation of Taiwan's Upper Secondary Education System: A Policy Analysis" (Hueih-Lirng Laih and Ian Westbury); (19) "The Internet and the Truth about Science: We Gave a Science War but Nobody Came" (George Meadows and Aimee Howley); (20) "Critical Evaluation for Education Reform" (Gisele A.

+++++ ED429093 Has Multi-page SFR---Level=1 +++++

Waters); and (21) "Boundary Breaking: An Emergent Model for Leadership Development" (Charles Webber and Jan Robertson). (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 429 093

VOLUME 6 - 1998

ISSN 1068-2341

Education Policy Analysis Archives

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL HAS
BEEN GRANTED BY

Glass

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

1

editor:
Gene V. Glass
Arizona State University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM029630

BEST COPY AVAILABLE

Volume 6 1998

EDUCATION POLICY ANALYSIS ARCHIVES

Article Pages
1 Dorn: The Political Legacy of School Accountability Systems 39
    Commentary 3
2 Glenn: Review of Arons's Short Route to Chaos 6
3 Arons: Response to Glenn 7
4 Gouvias: Selection in Europe--the Case of Greece 38
5 Zhao: Restructuring Chinese Higher Education 17
6 Tanner & Stone: School Improvement Policy--Site-Based Management 26
7 Akkari & Perez: Educational Research in Latin America 12
8 Broadhead & Howard: "The Art of Punishing": The Research
    Assessment Exercise 16
9 Slowinski: SOCRATES Invades Central Europe 30
10 Wilson: Educational Standards and the Problem of Error 304
11 Berg: Public Policy on Distance Learning in Higher Education 18
12 Dogan: Counseling in Turkey: Current Status and Future Challenges 14
13 Mehrens: Consequences of Assessment: What is the Evidence? 35
14 Stake: Some Comments on Assessment in U.S. Education 9
15 Hanish, Horan, Keen & Clark: A Note on the Empirical Futility
    of Labor-Intensive Scoring Permutations for Assessing
    Scholarly Productivity 12
16 Levin: Criticizing the Schools, Then and Now 13
17 Hackett & Carrigan: Performance Indicators 31
18 Laih & Westbury: Transformation of Taiwan's Upper
    Secondary Education System 31
19 Meadows & Howley: The Internet and the Truth about Science 13
20 Waters: Critical Evaluation for Education Reform 47
21 Webber and Robertson: Boundary Breaking: An Emergent
    Model for Leadership Development 30

EPAA • enter the archives • browse the abstracts • the editors • the edit board
• submit article • submit commentary • search • subscribe
volume: • 1 • 2 • 3 • 4 • 5 • 6 • 7

This article has been retrieved 4387 times since January 2, 1998.

Education Policy Analysis Archives

Volume 6 Number 1 **January 2, 1998** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal. Editor:
Gene V Glass Glass@ASU.EDU. College of Education
Arizona State University,Tempe AZ 85287-2411
Copyright 1998, the EDUCATION POLICY ANALYSIS ARCHIVES.Permission is hereby granted to copy any article provided that EDUCATION POLICY ANALYSIS ARCHIVES is credited and copies are not sold.

The Political Legacy of School Accountability Systems

Sherman Dorn
University of South Florida

Abstract

The recent battle reported from Washington about proposed national testing program does not tell the most important political story about high stakes tests. Politically popular school accountability systems in many states already revolve around statistical results of testing with high-stakes environments. The future of high stakes tests thus does not depend on what happens on Capitol Hill. Rather, the existence of tests depends largely on the political culture of published test results. Most critics of high-stakes testing do not talk about that culture, however. They typically focus on the *practice legacy* of testing, the ways in which testing creates perverse incentives against good teaching. More important may be the *political legacy*, or how testing defines legitimate discussion about school politics. The consequence of statistical accountability systems will be the narrowing of purpose for schools, impatience with reform, and the continuing erosion of political support for publicly funded schools. Dissent from the high-stakes accountability regime that has developed around standardized testing, including proposals for professionalism and performance assessment, commonly fails to consider these political legacies. Alternatives to standardized testing which do not also connect schooling with the public at large will not be politically viable.

Introduction

The short-term question about high-stakes testing is not whether it shall prevail but who shall control it. The president of the United States advocates the use of standardized testing developed by the federal government. (Note 1. Opens in separate browser window.) Conservatives who vigorously oppose nationalized curriculum and testing agree that testing should exist, but organized on a state and local level instead (see Diegmueller and Lawton 1996; Lawton 1997). The recent compromise between Rep. William Goodling and the White House left the long-term fate of a truly national testing program unresolved (Hoff 1997). Nonetheless, what is not at stake is the existence of high-stakes testing. Recent polling suggests that the idea of national testing is very popular (Rose, Gallup and Elam 1997), and that popularity reflects the past twenty years' growth of standardized testing. The debate over the control of testing takes for granted the existence of standardized testing because of its recent history. States for many years have been accumulating testing requirements which their legislatures, state officials, or local administrators have chosen. Despite considerable evidence that high-stakes testing distorts teaching and does not give very stable information about school performance, test results have become the dominant way states, politicians, and newspapers describe the performance of schools. Some have continued to note the problems of high-stakes standardized testing (e.g., Madaus 1991; McGill-Franzen and Allington 1993; Neill 1996; Noble and Smith 1994; Shepard 1991; Smith 1991; Smith and Rottenberg 1991; Wirth 1992: Chap. 7). Others try to accommodate some measure of standardized testing while building what they see as safeguards against obvious abuses. Still others (administrators in systems or schools with above-average test scores) use results as part of a marketing or public relations strategy. Few critics of high-stakes testing, however, have explicitly noted the way in which the public use of accountability systems shapes the politics of education writ large.

Statistical accountability systems are important because numbers have visible power in public debate. Anyone who listens to or reads politicians, journalists, and social critics will hear statistical references. Slowly over the last century, statistics have taken a prominent place in political culture. Whether the statistic is the official unemployment rate, poverty rates, poll results, or SAT scores, a specific number fills a niche in discussion. As Carol Weiss (1988: 168) wrote,

> The media report the proportion of the population that has been out of work for fifteen weeks or more, characteristics of high schools which have the highest drop-out rates, reasons given by voters for choosing candidates. These kinds of data become accessible and help to inform policy debates.

A number connotes objectivity or, at the very least, legitimacy. Because we perceive numbers and statistics as having a certain force on its face (just by being quantitative), we allow statistics to shape our perception of the world and the issues we perceive as important. They present selective information and thus center discussion around specific topics (silencing others). Nonetheless, we often yearn for the end of political uncertainty through statistics. Partisans in a conflict may heatedly argue that their methods are better, or their opponents' use of statistics is politically motivated, yet behind the veneer of cynicism lurks a desire for unquestionable statistics that will end debate. Maybe the official poverty line is arbitrary, but others have calculated alternative poverty estimates (Axinn and Stern 1988: 73-77; Ruggles 1990). The portrayal of a "rising tide of mediocrity" in schools was an alleged lie, but then the critics presented their own statistics as counter-evidence (Berliner and Biddle 1995; Bracey 1991, 1992, 1993, 1994, 1995a, 1996, 1997; National Commission on Excellence in Education 1983).

The production and presentation of statistics is part of the fabric of public debate, and public policy that involves the heavy use of statistics must consider the long-term consequences of that use. At least two such consequences are important, what I will call the *practice* and *political* legacies of statistics. The distinction between the two revolves around related but heuristically distinct issues:

- How do policies based on statistics shape practice?
- How do policies based on statistics shape future public policy debate?

The *practice legacy* of statistics is the nuts and bolts of how statistics shape government and private action. For example, the official U.S. consumer price index determines cost-of-living indices for Social Security, government pay schedules, and the behavior of many private organizations. Census population counts determine state representation in the U.S. House of Representatives and some federal spending patterns. This practice legacy can, by itself, engender vivid disagreement about statistical mechanisms. In 1997, several so-called deficit hawks suggested changing the calculation of the consumer price index to lower cost-of-living indices deliberately. While they claimed that the official inflation statistics misrepresented the "true" amount of inflation, reporters and groups such as the United Auto Workers clearly understood that the argument was not about the most accurate picture of inflation but was, in large part, about the practice legacy of inflation statistics for the U.S. federal budget, entitlement programs, and private company wages and benefits (e.g., "Will Washington Cut Our COLA?" 1997). Similarly, debate about the conduct of the decennial U.S. census in the past ten years has revolved not around accuracy but policy consequences. If, as some have proposed, the Bureau of the Census augments its population count with samples to measure undercounting and adjusts the official counts with the help of samples, the distribution of federal aid to cities and states as well as Congressional representation will change according to adjustment for undercounting. Politicians in jurisdictions with alleged undercounting have an interest in supporting such adjustment based on sampling because adjusted population counts would give their constituencies higher federal aid. Other politicians have an equally intense incentive in opposing the use of sampling to prevent the loss of federal aid (Mears 1997; Roush 1996). The practice legacy of statistics is an obvious consequence of tying statistics to public policy. The examples above show specific practice legacies, when statistics are mechanisms of what Paul Starr (1987: 55-57) calls "automatic pilots." They may be less obvious in the creation of systems of incentives, as some argue that high-stakes testing environments create. Whether the result is from explicit formulae or a consequence of incentives, a practice legacy is the influence of policy on short-term behavior.

What is less clear, but equally important, is the *political legacy* of statistics, the way that the use of statistics by itself shapes public debate. (Note 2. Uses second browser window.) Discussion about teenage pregnancy is a good example of how the existence and distribution of statistics shapes debate. In the late 1960s and early 1970s, as teenage birth rates were decreasing, the Alan Guttmacher Institute and others began publicizing estimates of teen fertility statistics to illustrate what they termed an epidemic of teenage pregnancy. The social construction of teen pregnancy as a *growing* problem contributed to political support for policies such as family planning and has been critical in debates over the consequences of family planning policies, even when the statistics were questionable (Vinovskis 1988). Feminism also contributed to changing attitudes towards family planning policies, but the paradox for social scientists is that demographic trends did not affect perceptions of the levels of teen pregnancy. Academic researchers on teen pregnancy have recognized the incongruity that the definition of teen pregnancy as a social problem coincided with a decrease in birth rates (e.g., Furstenberg 1991). Still, gross numbers (for example, total births to teen mothers) created the popular perception of a crisis. Statistics help define perceptions of social realities and possibilities. Starr (1987: 54) has noted,

> An average is not just a number; it often becomes a standard. . . . Many regularly reported social and economic indicators have instantly recognizable normative content. The numbers do not provide strictly factual information. Since the frameworks of normative judgment are so widely shared, the numbers are tantamount to a verdict.

The existence and frequent public reporting of teen pregnancy statistics by themselves created public debate that led to policies attempting to limit teen pregnancies. Much other public reporting of statistics likewise shapes public debate: Newspapers and broadcast news regularly report unemployment and inflation figures, crime rates, and school test scores.

The distinction between practice and political legacies of statistics is useful in explaining why accountability practices are so popular and what the potential consequences of the most commonly-discussed accountability systems might be in the long term for school politics. Most critics of high-stakes standardized testing point to the practice legacy, the way that high-stakes testing may narrow the focus of teaching and provide perverse incentives within schools and school systems. However, the political legacy is as important as, and in some important ways dovetails with, the practice legacy. High-stakes testing narrows how we judge schools as institutions and whose school success is important.

Moreover, opponents of high-stakes testing rarely consider the political legacy of proposed alternatives. The most prominent alternative vision of accountability revolves around the outdated model of ascendant professionalism. A consideration of accountability's political legacy would require different alternatives to high-stakes testing, ones that would cultivate deliberate political connections between schools and communities.

Table of Contents

- The Importance of Political Legacies
- The Popularity of School Accountability
- Unexamined Assumptions of Accountability
- The Political Costs of Accountability
- The Political Weaknesses of Professionalism
- The Ground We Stand on
- Where To Go
- References (uses a second browser window)

The Importance of Political Legacies

I choose the term *political legacy* for statistics because statistical systems constitute a special example of how public policy creates long-term consequences for public debate. Those who study government from a variety of disciplines recognize that public policies set in motion political dynamics that shape the contours (and sometimes define the limits) of accepted political debate. Two parts of the original Social Security Act of 1935, pension insurance and Aid to Dependent Children (the federal program most call welfare), demonstrate the way that policies can define the political landscape. The pension insurance part of Social Security is a universal program; anyone who pays into Social Security as a wage-earner (as well as a beneficiary defined by law) is eligible for payments when older. The universality of the Social Security pension has made its basic features unassailable politically. By contrast, federal welfare was a means-tested program. Only poor people (and not all poor people) were ever eligible for federally-supported welfare programs. Unlike Social Security pension insurance, welfare was politically vulnerable because of its means testing. Since most people would like to live long, they think of Social Security as an important safety net. But most people do not want to be poor and, as critically, may not think they ever will be poor enough to be on welfare. The universality of Social Security has protected it politically. Thus, when President Ronald Reagan suggested changing the pension program in the early 1980s, politicians rallied to support the system. However, without universality, federal welfare had a much less powerful base of support, and the Republican Congress and President Bill Clinton ended the federal welfare guarantee in 1996. The original outlines of the two programs shaped future debate over them (Skocpol 1991).

The different histories of school desegregation in the South and elsewhere since 1954 are also results of a political legacy. The fundamental paradox of desegregation is that the South (including border states) had the most integrated schools in the country by the late 1980s (Orfield 1993). Southern schools have been more integrated because of two policies vigorously pursued by white, racist politicians and officials before 1954: state laws mandating segregation and policies of school and government consolidation. Because state law and intentional acts by school officials were an obvious cause of school segregation, federal courts after 1954 had clear and convincing evidence of unconstitutional segregation in Southern systems and were willing to order far-reaching remedies in the late 1960s and early 1970s. In addition, Southern school systems are usually much larger than systems in many other states because of consistent success in consolidating school systems this century. For example, Mecklenburg County, North Carolina, has one school administration, so the suburbs of Charlotte are in the same school system as the city. In contrast, the suburbs of Boston are in school systems separate from the central city. Desegregation advocates in the South had two advantages stemming from consolidation. First, courts were more willing to order metropolitan desegregation plans in the South, after the *Milliken v. Bradley* (1974) decision required that judges find specific evidence of discriminatory intent to remedy metropolitan segregation in fragmented urban areas. Second, large systems made white flight more difficult. Because the South had both a history of state-directed discrimination and also large school systems, desegregation efforts in the region in the late 1960s and early 1970s were more vigorous and far-reaching than in the rest of the U.S (Douglas 1995; Orfield, Eaton, and the Harvard Project on School Desegregation 1996). The political legacy of statutory segregation and school consolidation made extensive desegregation more feasible in the South.

These stories, of government pension and welfare programs in one case and desegregation in the other, demonstrate the relationship between the structure of public policy and later political decision-making. To be sure, that influence is not one-way. A government is not an empty vessel easily manipulated by electoral and other political forces. Instead, government agencies have their own interests, and officials often act in their organizational interests (Balogh 1991b; Galambos 1970). Schools, like other public bodies, have their own professional and organization dynamics that mediate, rather than automatically reflect, outside influences. Thus, when we speak of a political legacy of school policies (including statistical systems), that legacy is part of a larger negotiation over the role of public schools. Two facets of that constant bargaining are particularly relevant to understanding the current school accountability regime: the limits of educators' professional authority and the local nature of schooling. First, as explained in the next paragraph, school administrators have tried to claim both bureaucratic autonomy and public acknowledgement of expertise involved in running schools. They have been far more successful in the former task than in the latter. In addition, schooling is a local, public service. Local political control of schools, and the close watch that one can theoretically keep over such institutions, may be one reason why school administrators garnered autonomy earlier in this century. One can thus view statistical accountability systems as one way to resolve the dilemma between granting autonomy and authority to educators and keeping them under some political control.

The political legacy of statistical accountability systems is important because support for publicly controlled schools is fragile. School administrators deliberately built a set of bureaucratic institutions in the early twentieth century to buffer themselves politically, in part by claiming the need for autonomy to exercise professional judgment and wield their expertise (Tyack 1974; Tyack and Hansot 1982). That autonomy, and the justification for publicly controlled schooling, has been on the wane since mid-century for several reasons. First, the civil rights movement targeted schools as one public institution that was treating poor and minority children unequally. The attack on school inequalities undermined support both from those who thought that inequality is morally wrong *and also* from those who had relied on state and local control of education to preserve bastions of private privilege (Kozol 1991). Second, the credibility of public institutions as a whole has deteriorated. In part, the Vietnam War and Watergate created a credibility gap between what public leaders said and what most citizens saw happening (Schell 1975); in addition, the internal politics of public agencies have damaged their ability to wield professional consensus as a political force (Balogh 1991a). Third, schools have been the target for half a century of accusations of ineffectiveness and soft standards. All of these events undermined the legitimacy of school administrators as autonomous professionals and public schools as worthy of financial and political support (Tyack and Hansot 1982). Privatization, through charter schools or vouchers, represents one potential result of declining support for school systems as publicly financed and controlled organizations. The political legacy of current educational reforms, including growing development of statistical accountability systems, will define in some measure the future debates about schooling.

Return to Table of Contents

The Popularity of School Accountability

The public judging of schools by test scores is relatively new in the United States. School statistics have existed since the late 19th century, and claims to objective measurement of student achievement from the turn of the 20th, but achievement scores have typically been only for internal consumption *within* school bureaucracies until recently. In the wave of school criticism after World War II, ideological debates over progressive education and the needs of the Cold War were the explicit points of conflict; statistical evaluations were invisible in the 1940s and 1950s debates over schooling (Ravitch 1983: 71-80, 228-32; Spring 1989: 10-33). The public debate over Scholastic Aptitude Test (SAT) score trends did not exist until the mid-1970s, even though the decline in mean scores began in the early 1960s. The *New York Times*, for example, did not start reporting SAT scores annually until 1976 (Maeroff 1976). No network news broadcasts between 1968 (when the Vanderbilt Television

News Archive began recording and indexing network news) and 1974 reported test scores as the substance of the story; the first networks to do so after 1967 were ABC and CBS on October 28, 1975. (Note 3. Uses second browser window.) The popular reporting of periodic student data, therefore, is of relatively recent vintage. One may consider statistics as one of many types of evidence and reasoning in public debate, such as the following list (meant to be an illustrative rather than a comprehensive typology):

Ideology
: Debates can focus on the purposes of schools and the perspectives offered in the curriculum or in teaching techniques. The attack on what progressivism had become by the 1940s is an example of ideological debate, as was the attack on outcome-based education in the early 1990s in Pennsylvania and elsewhere.

Representative Story
: Debates can center on real or apocryphal stories about education that represent the issue at hand. Anecdotes about high school graduates who cannot read (and the argued need for higher graduation standards) are an example of argumentation from representative story.

Statistics
: Debates over the quality of education in the 1980s, following the *Nation at Risk* report (National Commission on Excellence in Education 1983), are an example of discussion focused on statistics.

Direct Observation
: Debates can also focus on what individuals have seen, first-hand, in schools. I do not know of any national debate relying on directly observed evidence.

The self-evident explanation of the last statement suggests, in part, that we focus on statistics because having a "national discussion" based on personal, direct observation of schools is a contradiction in terms: we cannot each observe the nation's schools, and our judgment of "the nation's schools" will depend on second- or third-hand information. Still, most discussion of schools, and even school statistics, is local. Only thirteen network news broadcasts in the twenty-year period 1968-1987 reported statistical test score trends. (Note 3. Uses second brower window.) Most reporting on education, and most of what individuals hear and read from popular media sources, is still in local news broadcasts and local newspapers. Why, then, have local educational debates generally assumed the importance of statistics, something that makes more sense for a national debate?

The common use of statistical mechanisms to gauge school effectiveness, including the power of standardized test scores, owes its existence to the tension between the development of a national debate over education in the twentieth century and the continuation of local decision-making. The result is a set of themes which dominates discussion in cities and states across the country and that borrow much of their character and assumptions from the national debate. In many cities and towns, for example, newspapers and local news broadcasts describe similar issues such as discipline problems and whether high school graduates are ready for the workplace. Several changes in schooling since the early 19th century have encouraged a national debate. First, educational reformers have typically borrowed from each other's ideas, spreading them from region to region. Second, professional educators and muckraking journalists in the late 19th and early 20th century explicitly campaigned in nationally-distributed journals against school corruption and the decrepit conditions in urban schools, on the one hand, and for professional autonomy on the other. Their campaign nationalized the Progressive Era education debate. Third, administrative progressives (as David Tyack has termed them) were successful in creating standard institutional routines in the first half of the 20th century, so that many school experiences adults remember now *are* much more similar across the country than adult memories of childhood were 150 years ago. We thus have a common set of experiences nationally, making the terms of debate familiar. Finally, the nationalization of politics more generally after World War II encouraged the debate over Cold War schooling described earlier. The civil rights movement and desegregation consolidated that national framework for discussion.

Still, the national educational discussion is a layer on top of and filtering down through older, local politics of schooling. Localism has remained a powerful force. It has controlled the politics of local *and federal* educational programs. For example, Southern members of Congress were critical in supporting federal vocational education programs early in the century because the federal government allowed Southern states to distribute funds disproportionately to white vocational programs and create different curriculum programs by race. The result was that vocational education programs served to reinforce the Southern caste structure (Werum 1997). Traditional federal deference to state action also modified and limited Title VI of the Civil Rights Act of 1964, whose implementation still helped force school desegregation in the South (Orfield 1969). Opposition to federal intrusion has limited national action to the present, including President Clinton's desire for tests created and organized by the federal government. Politicians are willing for schools to buy textbooks from national publishers, accepting a tacit national curriculum (Miller 1997). Federal *government* decision-making, however, threatens more than local control of curriculum; it threatens local political networks and ways of doing business. Local political control of school policies and funding thus vie with the national debate. The result is frequently a set of variations on common practices, resulting in the illusion of local control in many school matters. Standardized testing and accountability systems are one example of that limited variation. States are free to choose commercial tests, develop their own, or not to engage in high-stakes testing at all. Today, however, most local school systems or states test children in the spring using multiple-choice tests with scores that schools can compare (using the publisher's data) against a norming population of children in the same grade. In the past decade, many states and local districts have added real consequences for the tests, including publicly releasing score data. The result is a patchwork of high-stakes testing that covers most of the nation. Despite theoretical local choice about standardized testing, one way of publicly judging schools has become dominant.

The emergence of contemporary school "accountability" dependent on test score results combined an existing set of practices (standardized testing) with the judgment of local schools within a national framework. Within a decade, public judgment of schools by test statistics became common, after the College Board publicized the decline in mean SAT scores, states began instituting minimum competency tests, and the National Commission on Excellence in Education published *A Nation at Risk* in 1983. Two historical perspectives underline the importance of understanding the political implications of school accountability systems.

- Accountability has turned the use of educational statistics upside-down. Statistics bolstered the claims of administrators to expertise early in this century, but politicians and popular news media now use statistics to judge school systems. This reversal shows the weakness of local school administrators in claiming professional authority. Autonomy within bureaucratic organization, not public respect of their expertise, is the primary power of school officials.
- The popularity of published test scores obscures alternative ways of judging schools. In less than twenty-five years, statistical accountability has become so ubiquitous that it appears inevitable. The change has been, in retrospect, both breathtaking and alarming in its speed. Political debate over the meaning of statistics has largely eclipsed other ways of describing what happens in classrooms.

The dominance of educational test scores today hides the fact that we did not have to use statistics as the dominant way of describing schools and their problems, and that in the past we have used many other means. Even when we evaluate local schools using nation-wide questions, we can use many sources of information. Assuming we must use primarily statistics is dangerous. We must remember that the evaluation of schools by test score statistics is one *among many* possible ways of seeing education through both national and local perspectives. Whether we made that choice consciously or wisely is a different question.

Return to Table of Contents

Unexamined Assumptions of Accountability

One consequence of public policy is the definition of legitimate debate and, by extension, what is not part of mainstream public discussion. Often, the assumed axioms underlying policies silence other relevant concerns (Fine 1991: 32-34). Despite more than twenty years of debate about the statistical performance of students in the U.S. and the proper direction for school reform, remarkably few voices in public have questioned the primary assumptions behind the move towards accountability. This silencing shows what we are avoiding when we speak glibly of a political consensus around school accountability. While we are agreeing to high-stakes testing, what uncomfortable issues are we not discussing? The broad political legacy of statistical accountability systems is the narrowing of legitimate topics for public debate. We do not often discuss the purpose of accountability or who will be making the key decisions to keep schools accountable.

Accountability for what purposes?

The dominant discussion of accountability leaves vague the goal of accountability mechanisms. The improvement of schools is an insufficient goal because accountability is fundamentally a political and not a technical process. Accountability has multiple meanings, in both a general sense and also the current sense in education of statistical judgment (Darling-Hammond and Ascher 1991). The apparent consensus for "accountability" hides the differences (and the conflicts) among the following meanings of statistical systems.

Judging public schools as institutions. One may use test score statistics to judge schools as a set of institutions. This sense of accountability (judging the worth of schools in general by test scores) is one of the most widely used tools in school politics. The annual release of average SAT scores in the late 1970s prepared the ground politically for the claim of declining school effectiveness made by the National Commission on Educational Excellence (1983). One political legacy of judging public schooling by test scores is the assumption that schooling is a monolithic entity that fails or succeeds as a single body. What this myth of a monolithic system hides is wide variations in schooling, especially between poor and wealthy schools (Kozol 1991). Another political legacy is that, after intense media focus on statistics that suggest poor schooling, citizens may face difficulty reconciling popular conceptions of failing schools with information gathered in other ways. Polls consistently show that parents' perceptions of their local schools are more positive than their perceptions of schooling nationwide (e.g., Rose, Gallup, and Elam 1997). In addition, private interests may subvert policies based on the gross judgment of schools. For example, some wealthy parents in one Michigan district deliberately pulled their children out of high-stakes standardized testing when they perceived that it might hurt their children (Johnston 1997). They may well have been willing to have high-stakes testing for "other people's children" (to borrow from Lisa Delpit's 1995 book title) but not theirs. This consequence is the educational equivalent of urban development NIMBY (Not in My Back Yard) syndrome.

Judging teachers and other educators. One may also justify accountability as a way to raise (or clarify) expectations and goals for teachers and administrators. An explicit part of accountability systems in the last few years has been the evaluation of teachers, principals, and other administrators. For example, the Tennessee Value-Added Assessment System, passed in 1992, originally mandated statistical measures of student gain as part of personnel evaluation (Educational Improvement Act of 1992). An earlier variant of judging teachers, schools, and school systems by comparative statistics was the U.S. Department of Education's "Wall Chart" instituted by Terrence Bell as an attempt to spur reform (Ginsburg, Noell, and Plisko 1988). This use of accountability, focusing on teachers and administrators, is the one most criticized as encouraging teaching to the test and "gaming" test results (Cannell 1989; Glass 1990; Madaus 1988, 1991; McGill-Franzen and Allington 1993; Merrow 1997; Shepard 1991; Smith 1991; Smith and Rottenberg 1991). The political legacy, however, may be even more harmful: By setting up a system based on the distrust of teachers, we make alternative ways of judging teachers and schools more difficult (Fisher 1996; Sizer 1992: 188-89).

Judging students. In many states and school systems, standardized tests have high stakes not only for educators but also for individuals students, as scores can be among the criteria for entrance to academic programs, grade promotion, or other real rewards and punishments in schooling. The use of tests to sort students U.S. began with monitorial schools in the early nineteenth century and admissions tests to early public high schools (Kaestle 1973; Labaree 1988; Reese 1995). More recently, the use of so-called minimum competency tests emerged in the late 1970s as a response to allegedly lowered standards of public schools (Bracey 1995b). The rationale of using tests to make students accountable is that, having test scores as a clear goal, students and schools would meet the expectations (Ravitch 1995). One potential legacy of such high stakes, however, is the rhetorical scapegoating of students. Calhoun (1973: 70-72) describes one purpose of testing in schools as displacing blame for ineffective teaching onto students. If a student fails a test, one may reason, the failure is the student's intelligence and lack of diligence. That consequence is already evident in many states with high-stakes testing. In Tennessee, for example, the teachers union pressed to exempt scores of students with disabilities from teacher value-added statistics ("Sanders model to measure 'value added'" 1991). One might presume that children with disabilities are those on whom we should *most* focus attention in evaluating teaching effectiveness. Yet teachers asked for the exclusion of scores because, the union argued, including such scores would be unfair to teachers. The displacement of blame for failed schooling onto students is a legacy of testing that existed well before high-stakes standardized testing, but accountability systems may exacerbate such tendencies (e.g., McGill-Franzen and Allington 1993; McGrew, Vanderwood, Thurlow, and Ysseldyke 1995; National Center on Educational Outcomes 1994).

Judging public policy. One might use standardized test scores (like other information) to evaluate public policies. The National Assessment of Educational Progress (NAEP) tests, begun in 1969, is theoretically a means for using non-high-stakes testing to evaluate public school policy with objective data. NAEP data is at the heart of some recent debate about school and student performance (see Berliner and Biddle 1995, 1996;

Stedman 1996a, 1996b). However, demands to use the NAEP to judge educators and students in high-stakes systems is threatening to compromise NAEP's use as a lower stakes way to gather information about student performance (Jones 1996; Koretz 1992a). One problem is the technical and fiscal demands of high-stakes versus low-stakes systems. In addition, however, is the ideological debate about the use of information. Can one maintain a low-stakes statistical system in the face of political pressures for high-stakes accountability?

Building organizations. In a broad sense, standardized testing supports the determination or control of curriculum content at the state and national levels. Some such as Ravitch (1995) explicitly advocate curriculum content standards and see teaching to the test as valid with appropriate testing and content. One consequence of statistical accountability, however, is the creation of new public and private organizations producing educational statistics. Publicly, states now have accountability or evaluation offices whose job is to provide the technical expertise in analyzing test data, and the federal government has the National Center for Educational Statistics, which contracts out NAEP as well as compiling and disseminating a wide variety of educational statistics. Private organizations supported by testing are the companies that write and sell tests or contract with agencies for the creation of specific tests. With each public release of test score statistics, popular news sources, politicians, administrators, and the public rely more on relatively anonymous technocrats to explain what is happening in schools. Other new professions this century, such as nuclear science, have also staked their claim to expertise on political factors (Balogh 1991a). The fact that this reliance on statisticians stems from political pressure for school reform usually escapes notice.

Marketing. Schools occasionally use student statistics as part of public marketing strategies, either to attract students who have choices (as in selective colleges) or to bolster public support. One of the largest metropolitan school systems in the country recently produced a pamphlet boldly titled, "Our Students' Test Scores Reflect Academic Achievement" (Hillsborough County Public Schools 1997). While one paragraph cautions that test scores are not the sole basis for evaluating students or schools, the rest of the pamphlet trumpets above-average achievement. Public relations was a strong motivation behind what Cannell (1989) called the "Lake Wobegon" effect of claiming high test scores in public reporting through the use of outdated norms. The use of accountability data for marketing is an open secret among administrators. As Dennie Wolf said in the John Merrow documentary *Testing . . . Testing . . . Testing* (1997), "Districts sell real estate based on test scores." With the decline of administrative authority described elsewhere in this article, superintendents have considerable interest in boasting about their systems using any tools at their command.

These varied purposes of accountability are not necessarily congruent. The use of test scores to bash public schools is not compatible with a nuanced debate over public policy, and students and teachers may have conflicts of interest when tests have high stakes for both. In addition to inconsistent purposes, the aims of accountability do not easily include other issues relevant to education: equity, the direction of curriculum, or the purposes of education more broadly in a changing world (Darling-Hammond 1992). One dominant assumption of accountability systems is that the goals of education are agreed upon and we need only establish a system to measure whether schools and students meet those goals. The creation of statistical accountability systems may freeze the assumption of a single purpose of statistical accountability into a framework for the politically accepted discussion in education for years hence.

Who keeps schools accountable?

A second unexamined assumption is that central bureaucracies and popular news media are the logical, natural places for holding schools accountable for performance. In most school testing regimes, central offices (at the state or local level) are responsible for the general logistics of testing and compiling results. Results at some level are then available to administrators, public boards of education, and media organizations. In many states and regions, newspapers publish test score statistics, often ranking schools or systems based on the scores. But who is *not* among the direct targets of test score dissemination is as important as who *is*.

Judges and advocates monitoring school system compliance in discrimination cases. Judges and advocates overseeing compliance with nondiscrimination orders (such as desegregation) generally are not intended users of "accountability" information. Despite promises by school systems to pay closer attention to achievement in desegregation cases, local systems have a very spotty record in demonstrating success after the end of desegregation orders. Orfield, Eaton, and the Harvard Project on School Desegregation (1996) has compiled evidence that, in several of the major cases this past decade, school districts released from desegregation monitoring by the courts not only experienced resegregation but growing achievement gaps between white and minority students. The new accountability system does not appear geared to keep systems accountable in this respect. Many advocates appointed to monitoring and advisory commissions have reported to Orfield and his associates that local systems have either denied information (such as disaggregated test scores) outright or made the gathering of data extremely difficult. In addition, the Supreme Court decision in *Missouri v. Jenkins* (1995) declared that district court judges should consider test scores as marginally important (at most) as a measure of compliance with racial equity requirements. The only major case where a court has continued to monitor standardized test scores as part of a major equity lawsuit has been in New Jersey, where the state's supreme court continues to criticize inequalities between the education offered children in the wealthiest and poorest systems of the state (*Abbott v. Burke* 1997). In the past five years, the court has broadened its focus from just monetary support of schools to include measurable outcomes. The New Jersey Supreme Court has been a lonely exception to the general rule, especially in the federal judiciary: Accountability does not appear to require even reasonably equitable outcomes.

Parents and the general public. Parents receive test scores of their children, but rarely do they or the general public have direct access to test score results *or their limitations.* Popular news sources (television, radio, and newspapers) mediate the transmission of information, often deleting information critical to understanding the limits of such data or transforming the statistics in ways either incomprehensible to readers or to create invalid statistical comparisons. The reporting of high-stakes test data by Nashville metropolitan newspapers form a case in point. Beginning in 1993, the state of Tennessee reported test results of schools and districts using a complex statistical system called the Tennessee Value Added Assessment System. The state's newspapers have quickly rushed to print school-by-school scores including rankings, even where schools many rankings apart had negligible differences in scores (in other words, when the rankings were unjustified by the statistics). For example, in 1996 the Nashville *Tennessean* transformed the value-added scores into percentile ranking, even though the technical documentation for value-added scores would not support such an interpretation (Bock and Wolfe 1996: Chaps. 5-6; Klausnitzer 1996; Tennessee Department of Education 1996). Why did the *Tennessean* transform value-added scores that were the result of a prior statistical manipulation, and why did the paper then rank schools? One reporter explained:

> We chose to report in percentile ranks because it helps people see how their school stacks up against the rest of the state, and because this information is not available anywhere else. It was calculated by *The Tennessean*... [because] we wanted to offer something unique. We also wanted to answer our readers' number one question about the test scores: How does my child's school compare to the other schools? (Lisa Green, e-mail to author, December 5, 1996)

In addition, the newspaper reported percentile rankings by tenths (for example, 50.1 instead of 50th percentile). The same reporter acknowledged that the newspaper staff did not consciously justify that apparent precision:

> There's really no need to report these numbers down to the tenth of a percentile. However, the programming for the site was written last year ... so the computer automatically included the decimal place, and we didn't think it was necessary to take it off. (Lisa Green, e-mail to author, December 5, 1996)

In this case, a metropolitan newspaper's desire to have "something unique" conflicted with its readership's interest in having clearly understandable information to interpret independently, or even information with a justifiable level of detail. Even if one assumes that the value-added scores are comprehensible, transforming those into percentile rankings was neither valid nor necessary for rankings (itself a method of reporting scores which the state's external evaluators recommended against). In no case did the newspaper note what the evaluators clearly stated: that school scores were unstable and could not be relied on for clear distinctions in performance (Bock and Wolfe 1996: Chap. 5-6). The dissemination of information through two intermediaries (the state government and news sources) in essence created one dominant way to analyze scores in the metropolitan Nashville area: how did schools "stack up" in competition with each other? The false precision in percentile rankings suggested that readers could rely on the numbers as rigorous, objective facts. The accuracy of newspaper reporting is also questionable; the *Tennessean* had to reprint its comparative tables in 1994 because of acknowledged gross errors in reporting ("How Midstate Schools Stack Up" 1994a, 1994b). While comparisons among schools may be appropriate in some ways, the presentation of school scores suggested a certainty which was incompatible either with the statistical calculations or the mediation of state agencies and newspapers in transmitting test scores.

Moreover, the dissemination and discussion of today's school accountability systems strip parents and the general public of control and ownership of information. In the case of Nashville, a reporter reduced parental evaluation of schools to examining rankings in a table, akin to sports league rankings (see Wilson 1996). One might contrast the typical method of disseminating accountability statistics with two alternative local methods of accountability: the "visiting committee" of town elders in the eighteenth and early nineteenth-century district schools, on the one hand, and the calculation of dropout statistics by a Hispanic activist organization in Chicago in the 1980s, on the other. In many district schools, a small committee of citizens held the power of hiring and firing over schoolteachers and could visit the school at any time (e.g., Cohen 1973: 407). Accountability in district schools was a rough-and-tumble affair, often unfair to teachers, but local citizens could form judgments in a simple way: watching classrooms. Independent gathering of data today is also possible. In the 1980s, Aspira, Inc., a Hispanic activist organization, suspected that official dropout statistics from the Chicago public schools were inaccurate or fraudulent and conducted its own research. Activists then used the independent statistics to help prod Chicago towards urban school reform (Hess 1991: 7-21; Kyle and Kantowicz 1991). In both cases, individuals at the local level produced and acted on their own judgments of schools. Reliance on centrally-calculated statistics in accountability systems often overrides local, independent judgment of schools.

The fundamental issue of control is directly connected to the purposes of accountability: Individuals in different roles would ask different questions of accountability mechanisms. Politicians might ask whether schools "measure up" to some standard (such as a national norm). Business leaders might ask about workplace-related skills and behavior. College faculty would want students to have some intellectual foundation. Parents might ask whether their children are getting enough individual attention. Who should be asking the hard questions about schools? The history of the Common Core of Data (a set of education data collected by the federal government since the early 1970s) illustrates the difficulties of creating an explicit consensus. Because of pressures within government, doubts about its utility and cost, and disagreements about what it should measure, the Common Core of Data for many years gathered relatively innocuous information in a history Janet Weiss and Judith Gruber (1987) described as "managed irrelevance." Of all the information used by the National Commission on Excellence in Education (1983) to lambaste the condition of schools, none came from the official federal education database (Weiss and Gruber 1987: 370). What we face is not an explicit consensus but a hidden one, never debated clearly, founded on the spread of standardized test scores. Statistical accountability systems suggest an objectivity and universality of coverage which is impossible. As Sizer (1995: 34) noted with regard to the debate about educational standards, "The word *system* has come up again; . . . Essentially, it implies a technocratic approach." We should not evade the political question of the purposes of schools through the production of statistics. The current penchant for statistical accountability systems diverts resources to a mechanism that hinders discussing the nuts and bolts of schooling. We hide behind the apparently objective notion of an accountability system.

Return to Table of Contents

The Political Costs of Accountability

The political legacy of statistical accountability systems is complex because of the different possible aims of (and justifications for) accountability and also because statistical systems will vary among different states and districts. Nonetheless, one can identify several broad patterns which stem at least in part from the proliferation of statistical accountability systems. Two legacies have seriously damaged our collective ability to have reasoned, broad discussion about the aims of schooling and reasonable public policy. Statistical judgment of school has narrowed the basis on which we judge schools and has also encouraged impatience with school reform.

Narrowed Judgment of Schools

Technocratic models of school reform threaten to turn accountability into a narrow, mechanistic discussion based on numbers far removed from the gritty reality of classrooms. Over the past twenty years, the dominant method of discussing the worth of schools in general has been the public reporting of aggregate standardized test score results. Popular news sources typically distort and oversimplify such findings (Berliner and Biddle 1995; Darling-Hammond 1992; Koretz 1992b; Koretz and Diebert 1993; Shepard 1991). The recent public debate over schools is not rich, reliant on multiple sources, or nuanced. Nor is the reliance on statistics inevitable in national discourse, despite recent history. Prior waves of reform, such as concerns about math and science education in the 1940s and 1950s (whether one agrees with their goals or not) did not need test score data as motivation or evidence (Ravitch 1983).

Test-score data and its use have pushed other issues to the margins. The aftermath of the 1983 report *A Nation at Risk* eclipsed two major policy initiatives of the first Reagan administration. The early 1980s saw dramatic cutbacks in the support of the federal government for state and local public schools. At the same time, social conservatives both in and out of the Reagan White House were arguing for the creation of vouchers to support parents sending their children to private schools. Neither of these issues, however, were part of the central discussion of education policy after the release of *A Nation at Risk*. The dominant discussion in popular news media revolved instead around declining test scores, the presumed responsibility of schools for national economic decline, and how to tighten academic standards (Berliner and Biddle 1995; Bracey 1995b). Few mentioned changes in the federal budget or privatization proposals, even though one was a concrete policy of the Reagan administration and the other was a radical proposal for changing the governance of schools. Ironically, the dominant discussion suppressed issues which concerned both liberals (upset at budget priorities) and social conservatives (wanting vouchers).

More recently, New Jersey Governor Christine Todd Whitman tried to argue that a standards-based accountability system alone could improve the state's schools. Her department of education responded to the state Supreme Court's call for equity with state-level achievement standards but no added resources, despite the state's history of vividly unequal funding among school systems. The argument by the executive branch was that standards, by themselves and despite existing funding inequities, would create school improvement. The assumption by Whitman is that test-based school accountability, as a technocratic mechanism with threatened sanctions, is sufficient to change schools, even schools with the worst records. The state court agreed with the governor in that New Jersey could have state-level standards but disagreed with the argument that funding was irrelevant. It then ordered the state to improve its funding of poor schools (once again) (*Abbott v. Burke* 1997). New Jersey is fortunate in having one branch of government able and willing to articulate a complex view of what school reform requires. In general, however, extending public discussion of schools beyond test-score statistics is difficult.

Impatience with Reform

On a political level, impatience with reform and the cyclical reporting of statistics encourages the dominant myth of contemporary educational politics, that schools continue to decline in quality. (Note 4. Uses second browser window.) That myth encourages a cynicism towards reform strategies. We should not be surprised that we have witnessed several "waves" of reforms since the regular publishing of SAT scores began in the 1970s. The mundane details of statistical accountability systems encourages fads. Without a concrete sense of what children and teachers should be or are doing, the public compares statistics against a set of arbitrary benchmarks.

On a practical level, statistical accountability produces both undue impatience with reform and laxity towards incompetence. The yearly reporting of test scores creates an artificial schedule for judging schools: Do they improve by the next set of annual tests? The periodic nature of reporting school statistics drives the disposal of reform writ large, because policy changes cannot change classroom practices on a deep and fundamental level or become institutionalized in a short time (Lipsky 1980; Tyack and Cuban 1995). Yet, paradoxically, the annual time-frame of standardized testing gives too *much* time for weak teachers to flounder without guidance or correction. Pinning personnel practices to annual testing may undermine the obligation of fellow teachers and administrators to keep a close eye on teachers without the necessary classroom skills. Principals may feel inclined to give poor teachers until the following cycle of annual tests to improve. For children, however, a year of being with an incompetent teacher can be extremely destructive. The problem is in part one of inappropriate time scales. Annual tests are too infrequent for appropriate guidance of instruction or evaluation of teaching, while they are too frequent to measure broader changes in schools.

In addition, standardized test accountability discourages the evaluation of what happens in the classroom. As long as a school or teacher has adequate test scores, what happens in the classroom is irrelevant. Similarly, poor test scores indicate needed change, no matter what happens in the classroom. The philosophy behind such practice-blind evaluation is putatively to give teachers autonomy. As the designer of one state's accountability system explained, accountability statistics allow teachers to make their own choices (Sanders and Horn 1994). Ultimately, however, this diminution of practice undermines teacher and school power, for several reasons. First, teachers do not usually have time to review and evaluate on their own a wide array of alternative teaching methods; they need support in selecting, adapting, and implementing different methods and curricula. Second, parents and other citizens *do* care about what happens in classrooms. Schools trying dramatic departures from normal practices face (sometimes very reasonable) criticism from parents even when the intent is to respond to the accountability system. Separating accountability from the sense of what a "real" school is (Tyack and Cuban 1995) is deceptive in the long run. It gives schools the following message: "Make your choices because we only care about test statistics. But we won't give you enough support to follow up on your choices, and in the end we will condemn your choices if they violate our ideas of what schools should be." One consequence of statistics-driven impatience is increased cynicism among teachers and administrators and their uncertainty about what the public really wants. Discussions isolated from what happens in schools may be politically alluring and attractive to popular news sources, but test scores drive a wedge between schools and the students and public they serve.

Parallels between Practice and Political Legacies

The political legacies of high-stakes statistical accountability systems parallel the practice legacies in two respects. First, narrowed political judgment of schools is the macropolitical equivalent of teaching to the test, a narrowing of the curriculum. Researchers have documented the tendency for teachers to narrow their focus to content and styles which they perceive will result in high test scores (Madaus 1988, 1991; Smith 1991; Smith and Rottenberg 1991; Shepard 1991). Relatively few teachers, faced with the onslaught of standardized testing, are willing to innovate. Meier (1997: 9) writes,

> The danger here is that we will cramp the needed innovations [in teaching] with over-ambitious accountability demands. Practical realism must prevail. Changes in the daily conduct of schooling . . . are hard, slow, and above all immensely time-consuming; they require qualities of trust and patience that we are not accustomed to.

High-stakes accountability is not a system that demonstrates trust in teacher's capacities. By signaling massive distrust, high-stakes testing instead provides low expectations for teachers (Sizer 1992: 110-13). Imagine the result of a thought experiment: the plight of John Dewey's University Lab School teachers under a high-stakes system. One might like to spend an extended time exploring history and science through the concrete example of textile manufacturing (Dewey 1899). In a modern accountability system, however, the state will test the children in March or April, with much of the test based on several dozen discrete skills. Whether the children can understand the role of textile mills in 19th century economic changes, or whether they can explain what principles allow a loom to work, is irrelevant to accountability systems based on standardized tests. Balancing such competing demands is extremely difficult. Teachers and schools who fight the pedagogical consequences of high-stakes testing are relatively unusual. Whether one agrees with the appropriateness of multidisciplinary teaching for some or all children, one cannot confuse the expectations of today's statistical accountability systems with expecting children to understand connections between what they see in life and academic disciplines. The latter is of a higher order of magnitude entirely. Relying on standardized tests and high-stakes production of test statistics is itself a dumbing-down of political debate and expectations for schools.

Similarly, impatience with reform and fad fetishes are the macropolitical equivalent of being impatient with children's progress. The aggregation of test score data often gives teachers and administrators incentives to exclude students whom they feel will harm test figures. Repeated reports of test scandals, the plea by teachers in Tennessee to exclude students with disabilities from their statistics, and variations in the proportion of students tested provide continuing evidence of the perverse incentives high-stakes testing provides (Glass 1990; Madaus 1988, 1991; McGill-Franzen and Allington 1993; McGrew et al. 1995; Smith 1991; Smith and Rottenberg 1991; Shepard 1991). These incentives perpetuate a dynamic of educational triage, wherein those who have the best chance to survive in life because of other circumstances also have the best opportunities to learn (Fuchs and Fuchs 1995; Sapon-Shavin 1993).

Return to Table of Contents

The Political Weaknesses of Professionalism

If accountability based on standardized tests encourages a narrow political discussion about education and impatience with schools, alternatives proposed by critics of standardized testing confront the same history that engendered statistical accountability. Dissenters from the accountability "consensus" exist, from longstanding standardized testing critics at FairTest (http://www.fairtest.org) to the Coalition for Essential Schools (http://www.ces.brown.edu) to Teachers College professor Linda Darling-Hammond and Arthur Wise, current president of the National Council for Accreditation of Teacher Education (NCATE). Each opposes the idea of motivating school reform by standardized testing. The proposed alternative methods of motivating better teaching include performance (sometimes called authentic) assessment of students, peer evaluation of teaching, and either creating a second tier of high-status teachers or restricting entry into a limited number of high-status positions within teaching. Advocacy of greater professional authority in education have generally focused on teacher education and preparation (e.g., Darling-Hammond, Wise, and Klein 1995; Holmes Group 1986; also see Labaree 1992), but includes accountability; for example, Wise has been concerned with the deskilling of teachers since *Legislated Learning* (1979). In general, the critics of standardized testing seek greater teacher autonomy and respect from the public, and in that way we might call professionalism the central value of the dissenters (e.g., Darling-Hammond 1988, Haefele 1992). Wise and Leibbrand (1993: 135) write that, "Hallmarks of a profession include mastery of a body of knowledge and skills that lay people do not possess, autonomy in practice, and autonomy in setting standards for the field." If teachers could successfully professionalize, Wise and others suggest, they would gain more respect from the public and earn the autonomy needed to improve schooling (e.g., Wise 1994). The logic of professionalism is very appealing with the explicit parallels to the professionalism of medicine (Starr 1982). It links mechanisms within schooling (who controls decision-making) to the public status of teachers and the politics of schools. Professionalism appears to be politically astute.

Professionalism, however, is not likely to be a successful gambit in schooling, for several reasons. Most importantly, professional ideology is politically unpalatable in the late twentieth century. Trying to use professionalism misunderstands the historical context for the ideology of expertise and its widespread (political) success a century ago. Professionalism in the form of high-status, science-based occupations like medicine and engineering was one response to the chaos of industrialization and changing class structure (Wiebe 1967). Its early proponents argued that the complexities of modern life required technical expertise to solve public policy and practical problems. However, professions include more than high-status jobs, with occupations as diverse as architecture and craft work like plumbing. A profession typically involves three dimensions: a claim to specialized expertise, some informal or formal credentialing to control entry into the occupation, and autonomy on the job (Friedson 1984). Classroom teaching falls partway among all three dimensions. Classroom teaching does involve some skills that few could walk in off the street with, but the general public has far more knowledge of what happens in classrooms (and is more willing to make second judgments of teaching) than fields like surgery. Long-term teaching requires credentials, but many school systems hire uncredentialed personnel on an emergency basis. Finally, public schools operate as loosely coupled organizations (Weick 1976): Most teachers can shut their doors in the face of some supervisory directives, but material conditions (such as the textbooks available) circumscribe their autonomy on the job, and they face other demands they cannot ignore, such as the official curriculum and standardized tests. We should see the ideology of professionalism thus as attempting to emulate a relatively small slice of all occupations with professional traits rather than, as is typically assumed, making teaching a "real" profession. Teaching already is a real profession, though one with less claim to specialized expertise and less autonomy than advocates of teacher professionalism would want.

Professionalism theories today appeal to an outdated ideal of insularity and ascendant authority. The worst excesses of school bureaucracies today stem from *successful* professionalism, albeit not in the classroom. Superintendents at the turn of the century argued that schools needed to be away from political battles that would harm the integrity of school systems. Creating an autonomous professional unit (a central school office) would improve administrative efficiency and rid schools of corruption (Tyack 1974; Tyack and Hansot 1982). Their success accelerated the bureaucratization of urban school systems. Today, however, professionalism is no longer unquestioned. School administration has credentialism and relative autonomy on the job, but not as much claim to specialized expertise as sixty or seventy years ago. Not only are North Americans far more skeptical of professional authority than fifty years ago (as discussed earlier), but capital mobility is impinging on professional authority in a wide range of fields. The parallels made between teacher professionalism and medical professionalism is jarring. One cannot today call medicine an autonomous profession when doctors are complaining that clerical workers and financial officers in health maintenance organizations are limiting their clinical decision-making (Bodenheimer 1996).

In addition to ignoring the historical decline of professionalism, arguments for advancing teacher professionalism undermines democratic control of schools. As Strike (1990: 362) noted, "Professionalism is nondemocratic in that it appeals to political values other than those of popular sovereignty to legitimate its authority." Peer review of teaching (e.g., Haefele 1992) is a case in point. Civil rights activists may not want teachers to have virtually unlimited autonomy in the classroom. Bob

Peterson (1997: 4) explained, "A potential problem with the strictly professional union approach [to accountability] . . . in many urban districts has distinct racial overtones. Is peer evaluation the exclusive province of teachers and administrators or should parents and community members play a role?" Especially as the teaching force's demographics diverges from those of students and parents (Justiz and Kameen 1988), relying on professional-only evaluation may insult parents of a school who expect a role in school governance. Having an expertise-based evaluation system conflicts with U.S. traditions of democratic control, upon which civil rights activists have based advocacy of school governance councils. Some critics of standardized testing, such as Wilson (1996), point to British school inspections as an alternative to statistical accountability. The heart of the British inspection system, however, was until recently a self-perpetuating corporate body selected by and from experienced teachers. One may (as Wilson did) use school inspection to point out the problems in high-stakes accountability. One may not, however, successfully import the insular assumptions of professionalism to late 20th United States public schooling.

Professionalism is the dominant alternative to standardized-test-based accountability. Other critics of standardized testing-based accountability may not be as explicit as Wise in their advocacy of professionalism, and they may not agree with his proposals to limit entry into high-status positions in teaching. Still, they argue for more decision-making power in the classroom and school and see the bureaucratization and centralization of authority as one of the reasons why standardized testing is flawed. Thus, Kenneth Peterson (1995: 4) argues that one of the key principles in teacher evaluation should be to "place the teacher at the center of evaluation activity." In that respect, the professionalism label is a useful heuristic device for understanding opposition to standardized testing. Despite its intriguing hypothesis (that status and autonomy are the key to educational reform), professionalism is unlikely to supplant high-stakes accountability because it is politically untenable.

Moreover, professionalism addresses primarily concerns inside schools (autonomy of teachers). Publicly, professionalism only changes the superficial aspect of teacher status, not the public dissatisfaction and disconnection which schools face more broadly. Several historical changes have fragmented what is supposedly a common public commitment to education. The aging of the population since the height of the baby boom has shrunk the political power of parents. In addition, the civil rights movement and a political coalition of fundamentalist Protestant organizations have stripped school officials of any broad political consensus. Finally, the fragmentation of urban politics and suburban growth has encouraged continued racial and class segregation (albeit in new forms), making common interests in broad school policies difficult (Katznelson and Weir 1985). While I doubt professionalism's proponents would ever claim that it is a panacea, they have nonetheless pinned their hopes for dramatic school reform on a model that would not solve the major problems of school politics today.

Return to Table of Contents

The Ground We Stand on

Like the expansion of Israeli settlements in occupied territories, the continuing spread of standardized testing has created "facts on the ground" which have transformed both schools and the politics of education. To ignore the educational landscape around us, or to wish it would go away, is unproductive. Those who disagree with the assumptions of high-stakes, testing-based accountability must acknowledge that standardized testing is likely to become even more prominent in the short-term. This understanding should not prevent advocates from fighting the trend where possible. Local victories against high-stakes testing are important both to the children involved and also as a standing alternative to technocratic accountability. Nevertheless, we should see clearly what is and is not possible in the near-term future.

The Future Growth of Standardized Testing

Standardized testing connected with high-stakes accountability systems is likely to become more prominent in the next five years in the majority of states. The Education Commission of the States (1997) recently reported that almost half of all states have implemented or are planning public accountability systems using statistical measures. Some additional states may use the national tests advocated by President Clinton (if the tests exist). Some like Tennessee will design their own accountability mechanisms. Others like New Jersey will create a set of content standards with the promise of new tests and accountability tied to the content standards. The federal government and states will then spend millions of dollars developing tests, field-testing them, and supporting their use. In the meantime, popular news sources will continue to report annually the average SAT scores and tests currently used in local jurisdictions. Within five to ten years, some states will begin the mandated use of exams replacing or supplanting current off-the-shelf commercial tests.

Moreover, the political debate over tests is likely to center around the federal relationship between Washington and the states or (with privatization) public oversight of private schooling. For the duration of President Clinton's term, the administration is likely to support national tests, and governors who dissent (like Virginia's outgoing Governor George Allen) will do so not because they disagree with high-stakes tests but because they wish states to design their own independent standards. If federal courts, using *Agostini v. Felton* (1997), allow tuition voucher programs to proceed, state legislatures may contemplate mandatory use of high-stakes testing for private schools accepting public funds. The debate would then shift to public control of private educational institutions. A vision of the future debate may be *Ohio Association of Independent Schools v. Goff* (1996), in which a federal appeals-court panel concluded that Ohio's requirement to test private school students was constitutional. Those who disagree with all high-stakes testing will be at the margins of debate in the near future, except where they make alliances with others (as in the Congressional fight over national tests).

Limits on High-Stakes Testing

High-stakes testing has some significant weaknesses, despite the near-term growth we can expect. Some of the same dynamics which have limited the accountability use of performance-based, open-form testing will also shape standardized testing. Simply put, developing tests is expensive. The Tennessee legislature recently delayed the implementation of new subject tests for high school students to use in the value-added statistical system because, according to the bill's sponsor, the state could not afford the $10 million development cost (Educational Improvement Act Amendments 1997; Finn 1997). In addition, political adversaries may well use the management and pedagogical problems of new testing and accountability systems as a pawn in broader partisan battles. California's recent educational history is a case in point. Questions about the utility and propriety of performance-based tests combined with the expense of development and testing to kill the California Learning Assessment System. The governor, state superintendent, and legislature at the time were at odds over the purpose of the system, and that political conflict fed a controversy started by conservative critics over the ideological content of the tests, dooming the largest experiment in performance-based accountability to date (Kirst 1996; McDonnell 1997). Observers of merit pay have noted that political dynamics involving fairness and incentives to cheat typically kill merit pay systems (e.g., Glass 1990). The same may happen to the next generation of high-stakes accountability.

Contraction of the Meaning of "Public"

Despite the weaknesses of high-stakes testing, the short-term consequence of more standardized testing may be intensified criticism of public schooling and cynicism about the purposes of public educational systems. Schools need to be "public" in the sense of public involvement and political commitment (Fine 1991: Chap. 9; Katz 1992). However, the ranking of schools and teachers is inherently a zero-sum game, and not everyone can be above-average. Seeing school performance in such terms, divorced from classroom practice and public policy, makes both meaningful praise *and* criticism of schools very difficult. Moreover, the constant reinforcement of the myth of declining school performance will continue the erosion of support for the good schools that exist and make intense discussion of the needs of children more difficult.

Return to Table of Contents

Where To Go

Some alternative models of accountability may reverse the destructive tendencies of statistical accountability systems, both in political and practice terms. Reconstructing public education in its best sense (schooling *for* children, their families, and the public) requires connecting schools in a meaningful and explicitly political way with broader communities. In the same way that the development of the Central Park East elementary and secondary schools under Deborah Meier's leadership required both bureaucratic support and political connections to survive and thrive (Fliegel 1993; Meier 1995), so other schools and school critics dissenting from the current accountability trend must craft an alternative support structure, both within and extending beyond public schooling. Sizer (1992) argues for opening up schools to external evaluation for pedagogical reasons, to keep teachers in touch with reasonable expectations of what students should do. In addition, allowing friendly critics into schools serves an explicitly political purpose, giving community members a concrete sense of what happens in schools. No statistics can substitute for the type of immediate contact such external evaluation provides.

Permitting external evaluation is difficult today. Allowing strangers into schools is threatening because it erodes, at least on a symbolic level, the commitment to professional autonomy which administrators have maintained for almost one hundred years. In practical terms, it requires balancing the legitimate needs of teachers for enough time to plan and try out ideas against the interests of parents and the public to know what is happening in schools. In systems where many teachers may be from ethnic and racial groups different from their students, the tension between teachers and parents may be real, and letting parents into evaluation may be politically tricky (B. Peterson 1997). Yet educators must acknowledge the need to move beyond professionalism as the primary route to support for public schools. Isolating the workings of schools from the public has done teachers and administrators a disservice in the long term as professionalism has declined as a successful route to status and autonomy.

External community evaluation is not the only conceivable way of crafting alternatives to high-stakes standardized test accountability. Others might meet the same needs (e.g., Bernauer and Cress 1997). Common to solving the political problems of accountability are the following three requirements:

- Accountability should encourage deeper discussion of educational problems. Student performance should be the starting point of educational politics, not an occasion for political opportunism or crude comparisons. Statistical accountability, with the centralization of statistical production and dissemination through popular news sources, encourages oversimplification rather than a more extensive public discussion.

- Accountability should connect student performance with classroom practice. Statistical accountability, with the abstraction of student performance into numbers without context, removes classroom practices from the discussion of educational reform.

- Accountability should make the interests of all children common.

> This sense of commonality is the best meaning of "public" in public schooling. Statistical accountability systems intensify educational triage, encouraging schools to isolate and devote fewer resources to students whom schools judge as difficult to teach. Politically, statistical accountability systems divide the interests of schools and communities through competition for prestige and resources.

No one should pretend that accountability is without conflict or unproblematic. We should face those conflicts and issues directly, however, instead of hiding behind existing standardized testing. Some parents and others may well see statistical comparisons as a primary way for them to gauge school programs and children's education, or as a way to advance specific interests. For example, parents of students with disabilities and disability advocates face real quandaries over accountability. On the one hand, high-stakes testing has created incentives for segregating students (McGill-Franzen and Allington 1993). On the other hand, the national rhetoric emphasizing achievement for *all* students has provided a lever to criticize the omission of students with disabilities from assessment systems, to craft new federal law encouraging inclusion in assessment, and to create guidelines for state officials seeking to change assessment practices (Thurlow, Elliott, Ysseldyke, and Erickson 1996; also visit the National Center on Educational Outcomes site at http://www.coled.umn.edu/NCEO/). This dilemma is rooted in the tension between wanting to protect students with disabilities from the deleterious consequences of high-stakes testing and yet also wanting whatever accountability systems exist to pay attention to their interests. Those criticizing statistical accountability systems must understand this and similar dilemmas of parents and advocates. Changing attitudes and assumptions, while protecting what many see as important in statistical accountability, requires modeling of worthwhile alternatives and small-scale demonstrations that are explicitly political. Over time, if not immediately, schools need a plausible, fair way to evaluate school improvement. With enough local models of alternative accountability, then perhaps the dynamics of educational politics at state and national levels can change to become broader, connect with classroom practices, and require more than sound bites. Without those concrete examples, however, the domination of crude statistical evaluation of schools will continue, to the detriment of schools, children, their families, and the public.

Return to Table of Contents

Acknowledgements

I am indebted to Gene Glass and anonymous reviewers at *Educational Policy Analysis Archives*, Brian Balogh, Douglas Fuchs, and the students in my masters history of education class for contributions to my thinking about this subject. I am solely responsible, of course, for the interpretations here and the accuracy of detail.

Notes

1. I mean by standardized tests those administered in whole-group settings with quantifiable results. These include multiple-choice tests and also performance-based tests whose results are reportable in quantifiable terms. Thus, Advanced Placement exams conducted by Educational Testing Service are standardized tests for the purposes of this article because, even though parts of the test are performance-based (such as essays), the essays are scored by a quantifiable rubric system and the whole test reported on the company's 1-5 scale for such tests. Moreover, reporting scores by numbers allows the simplified public discussion which is my focus here. For an introduction to Lauren Resnick's advocacy of measurement-driven reform, see Simmons and Resnick (1993). For issues involved in Kentucky and Arizona (respectively), see Jones and Whitford (1997) and Noble and Smith (1994).

2. An anonymous reviewer noted that the line between practice and political legacies is fuzzy. In many ways, the debates over census undercounting and the consumer price index are also debates about the political rhetoric of the reapportionment process and future support for government entitlement programs. Nonetheless, the distinction between the two legacies is a useful heuristic device for explaining why the literature on perverse incentives of high-stakes testing does not address the critical issue of school politics.

3. According to the Vanderbilt Television News Archives, the following broadcasts discussed standardized test score levels between 1968 and 1987: October 28, 1975 (ABC, CBS); November 17, 1975 (CBS); August 23, 1977 (ABC, CBS); August 24, 1977 (ABC, commentary); September 1, 1977 (CBS, commentary); September 21, 1982 (CBS); September 19, 1984 (ABC); January 9, 1985 (NBC); January 26, 1985 (CBS, NBC); September 22, 1987 (NBC). The search terms included "standardized and test*," "test and scor*," "SAT and scor*," and "SAT and (college or scholastic)." Excluded from this list are stories about the alleged discriminatory nature of tests.

4. I agree with Stedman (1996a, 1996b) that schools are not as good as they should be. Those problems do not mean the myth of declining quality is true: schools have been inconsistent and too often mediocre for many years.

References

Note: Links here use a second browser window.

Abbott v. Burke. 1997. 149 N.J. 145. Retrieved September 27, 1997 from the World Wide Web: http://www-camlaw.rutgers.edu/decisions/supreme/m-622-96.opn.html

Agostini v. Felton. 1997. 117 U.S. 1997. Retrieved September 27, 1997 from the

World Wide Web: http://supct.law.cornell.edu/supct/html/96-552.ZS.html

Axinn, June, and Mark J. Stern. 1988. *Dependency and poverty: Old problems in a new world*. Lexington, Mass: Lexington Books.

Balogh, Brian. 1991a. *Chain reaction: Expert debate and public participation in American commercial nuclear power, 1945-1975*. New York: Cambridge University Press.

Balogh, Brian. 1991b. Reorganizing the organizational synthesis: Reconsidering modern American federal-professional relations. *Studies in American Political Development* 5 (1): 119-172.

Berliner, David C., and Bruce J. Biddle. 1995. *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. Reading, Mass.: Addison-Wesley.

Berliner, David C., and Bruce J. Biddle. 1996. Making molehills out of molehills: Reply to Lawrence Stedman's review of *The manufactured crisis*. *Educational Policy Analysis Archives* 4 (February 26). Retrieved September 27, 1997, from the World Wide Web: http://olam.ed.asu.edu/epaa/v4n3.html

Bernauer, James A., and Katherine Cress. 1997. How school communities can help redefine accountability assessment. *Phi Delta Kappan* 79 (September): 71-75.

Bock, R. Darrell, and Richard Wolfe. 1996. *Audit and review of the Tennessee Value-Added Assessment System (TVAAS): Final report*. Nashville, Tenn.: Tennessee Comptroller of the Treasury.

Bodenheimer, Thomas. 1996. HMO backlash -- righteous or reactionary? *New England Journal of Medicine* 335: 1601-1604.

Bracey, Gerald W. 1991. Why can't they be like we were? *Phi Delta Kappan* 73: 104-17.

Bracey, Gerald W. 1992. The second Bracey report on the condition of public education. *Phi Delta Kappan* 74: 104-8, 110-17.

Bracey, Gerald W. 1993. The third Bracey report on the condition of public education. *Phi Delta Kappan* 75: 104-12, 114-18.

Bracey, Gerald W. 1994. The fourth Bracey report on the condition of public education. *Phi Delta Kappan* 76: 115-27.

Bracey, Gerald W. 1995a. The fifth Bracey report on the condition of public education. *Phi Delta Kappan* 77: 149-60.

Bracey, Gerald W. 1995b. *Final exam: A study of the perpetual scrutiny of American education*.

Bracey, Gerald W. 1996. The sixth Bracey report on the condition of public education. *Phi Delta Kappan* 78: 127-38.

Bracey, Gerald W. 1997. The seventh Bracey report on the condition of public education. *Phi Delta Kappan* 79: 120-36.

Calhoun, Daniel Hovey. 1973. *The intelligence of a people*. Princeton, N.J.: Princeton University Press.

Cannell, John Jacob. 1989. *How public educators cheat on standardized achievement tests: The "Lake Wobegon" report*. ERIC Reproduction Document No. ED 314 454.

Cohen, Sol, ed. 1973. *Education in the United States: A documentary history*. New York: Random House.

Darling-Hammond, Linda. 1988. The futures of teaching. *Educational Leadership* 46 (November): 4-10.

Darling-Hammond, Linda. 1992. Educational indicators and enlightened policy. *Educational Policy* 6 (September): 235-65.

Darling-Hammond, Linda, and Carol Ascher. 1991. *Creating accountability in the big city school systems*. Urban Diversity Series No. 102. New York: National Center for Restructuring Education, Schools, and Teaching. ERIC Reproduction Document No. ED 334 339.

Darling-Hammond, Linda, Arthur E. Wise, and Stephen P. Klein. 1995. *A license to teach: Building a profession for 21st-century schools*. Boulder, Colo.: Westview Press.

Delpit, Lisa. 1995. *Other people's children: Cultural conflict in the classroom*. New York: The New Press.

Diegmueller, Karen, and Millicent Lawton. 1996. The road not taken. *Education Week* (April 24). Retrieved September 27, 1997, from the World Wide Web: http://www.edweek.com

Douglas, Davison M. 1995. *Reading, writing, and race: The desegregation of the Charlotte schools*. Chapel Hill, N.C.: University of North Carolina Press.

Education Commission of the States. 1997. *Accountability: State Policies*. Denver: Education Commission of the States. Retrieved September 27, 1997 from the World Wide Web: http://www.ecs.org/ecs/24aa.htm

Educational Improvement Act. 1992. Tennessee Acts Chapter 353.

Educational Improvement Acts Amendments. 1997. Tenn. Acts Chapter 434. Retrieved September 27, 1997 from the World Wide Web: http://www.legislature.state.tn.us/bills/100gahtm/100_chap/pubc0434.htm

Fine, Michelle. 1991. *Framing dropouts: Notes on the politics of an urban public high school*. Albany, N.Y.: State University of New York Press.

Finn, Michael. 1997. State school systems bill passes Senate. *Chattanooga Free Press* (May 31).

Fisher, Thomas H. 1996. *A review and analysis of the Tennessee Value-Added Assessment System: Part II*. Nashville, Tenn.: Comptroller of the Treasury.

Fliegel, Sy. 1993. *Miracle in East Harlem: The fight for choice in public education*. New York: Times Books.

Friedson, Eliot. 1984. Are professions necessary? In *The authority of experts*, edited by Thomas L. Haskell (pp. 3-27). Bloomington, Ind.: Indiana University Press.

Fuchs, Douglas, and Lynn S. Fuchs. 1995. Special education can work. In *Issues in educational placement: Students with emotional and behavioral disorders* (pp. 363-77), edited by James M. Kauffman, John W. Lloyd, Daniel P. Hallahan, and Terry A. Astuto. Hillsdale, N.J.: Erlbaum.

Furstenberg, Frank. 1991. As the pendulum swings: Teenage childbearing and social concern. *Family Relations* 40: 127-38.

Galambos, Louis. 1970. The emerging organizational synthesis in modern American history. *Business History Review* 44: 279-290.

Ginsburg, Alan L, Jay Noell, and Valene White Plisko. 1988. Lessons from the wall chart. *Educational Evaluation and Policy Analysis* 10 (Spring): 1-12.

Glass, Gene V 1990. Using student test scores to evaluate teachers. In *The new handbook of teacher evaluation* (pp. 229-40), edited by Jason Millman and Linda Darling-Hammond. Newbury Park, Calif.: Sage Publications.

Haefele, Donald L. 1992. Evaluating teachers: An alternative model. *Journal of Personnel Evaluation in Education* 5: 335-45.

Hess, G. Alfred. 1991. *School restructuring, Chicago style*. Newbury Park, Calif.: Corwin Press.

Hillsborough County Public Schools. 1997. *Our students' test scores reflect academic achievement*. Tampa, Fla.: Office of Communications and Governmental Relations.

Hoff, David J. 1997. White House, GOP craft agreement on testing. *Education Week* (November 12): 1, 23.

Holmes Group. 1986. *Tomorrow's teachers*. East Lansing, Mich.: Holmes Group.

How midstate schools stack up. 1994a. Nashville *Tennessean* (October 15): 4A.

How midstate schools stack up. 1994b. Nashville *Tennessean* (October 16): 6A-7A.

Johnston, Robert C. 1997. Just saying no. *Education Week* (April 9). Retrieved September 27, 1997, from the World Wide Web: http://www.edweek.com

Jones, Ken, and Betty Lou Whitford. 1997. Kentucky's conflicting reform principles: High-stakes school accountability and student performance assessment. *Phi Delta Kappan* 79: 276-81.

Jones, Lyle V. 1996. A history of the National Assessment of Educational Progress and some questions about its future. *Educational Researcher* 25 (October): 15-22.

Justiz, Manuel J., and Marilyn C. Kameen. 1988. Increasing the representation of minorities in the teaching profession. *Peabody Journal of Education* 66 (Fall): 91-100.

Kaestle, Carl F. 1973. *Joseph Lancaster and the monitorial school system: A documentary history*. New York: Teachers College Press.

Katz, Michael B. 1992. Chicago school reform as history. *Teachers College Record* 94: 56-72.

Katznelson, Ira, and Margaret Weir. 1985. *Schooling for all: Class, race, and the decline of the democratic ideal*. New York: Basic Books.

Kirst, Michael W., and Christopher Mazzeo. 1996. The rise, fall, and rise of state assessment in California, 1993-1996. Paper presented at the Annual Meeting of the American Educational Research Association. ERIC Reproduction Document Number ED 397 133.

Klausnitzer, Dorris. 1996. Pupils need practice on that first "R." Nashville *Tennessean* (November 15): 1A, 8A-10A. The *Tennessean*'s statistical analysis retrieved September 27, 1997 from the World Wide Web: http://www.tennessean.com/schools/ (defunct link)

Koretz, Daniel. 1992a. NAEP and national testing: Issues and implications for educators. *NASSP Bulletin* 76 (September): 30-40.

Koretz, Daniel. 1992b. What happened to test scores, and why? *Educational Measurement: Issues and Practice* 11 (Winter): 7-11.

Koretz, Daniel, and Edward Diebert. 1993. *Interpretations of National Assessment of Educational Progress (NAEP) anchor points and achievement levels by the print media in 1991*. Santa Monica, Calif.: The RAND Corporation. ERIC Document Reproduction No. ED 367 683.

Kozol, Jonathan. 1991. *Savage inequalities: Children in America's schools*. New York: Crown Publishers.

Kyle, Charles, and Edward Kantowicz. 1991. Bogus statistics. *Latino Studies Journal* 2 (May): 34-52.

Labaree, David F. 1988. *The making of an American high school: The credentials market and the Central High School of Philadelphia, 1838-1939*. New Haven, Conn.: Yale University Press.

Labaree, David F. 1992. Power, knowledge, and the rationalization of teaching: A genealogy of the movement to professionalize teaching. *Harvard Educational Review* 62: 123-54.

Lawton, Millicent. 1997. Riley delays national tests' development. Education Week (October 1). Retrieved October 2, 1997, from the World Wide Web: http://www.edweek.com

Lipsky, Michael. 1980. *Street-level bureaucracy: Dilemmas of the individual in public service*. New York: Russell Sage Foundation.

Madaus, George F. 1988. The distortion of teaching and testing: high-stakes testing and instruction. *Peabody Journal of Education* 65 (Spring): 29-46.

Madaus, George F. 1991. The effects of important tests on students. *Phi Delta Kappan* 73: 226-31.

Maeroff, Gene. 1976. Aptitude test lag is puzzling experts. *New York Times* (September 12): 17.

McDonnell, Lorraine M. 1997. *The politics of state testing: Implementing new student assessments*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Technical Report 424. Los Angeles, Calif.: CRESST. Retrieved November 14, 1997, from the World Wide Web: http://cresst96.cse.ucla.edu/Reports/tech424.pdf

McGill-Franzen, Anne, and Richard L. Allington. 1993. Flunk'em or get them classified: The contamination of primary grade accountability data. *Educational Researcher* 22 (January-February): 19-22.

McGrew, Kevin S., Mike L. Vanderwood, Martha L. Thurlow, and James E. Ysseldyke. 1995. Why we can't say much about the status of students with disabilities during education reform. NCEO Synthesis Report 21. Minneapolis, Minn.: National Center on Educational Outcomes. ERIC Reproduction Document Number ED 396 475.

Mears, Walter R. 1997. Infighting breaks out over census. CNN News (October 2). Retrieved October 2, 1997, from the World Wide Web: http://www.allpolitics.com/1997/10/02/ap/census (defunct link)

Meier, Deborah. 1995. *The power of their ideas: Lessons for America from a small school in Harlem*. Boston: Beacon Press.

Meier, Deborah. 1997. How our schools could be: Standards, top-down mandates, and grass-roots communities. *Rethinking Schools* 11 (Summer): 8-9.

Merrow, John. 1997. *Testing . . . testing . . . testing*. Public Broadcasting System. Script retrieved September 27, 1997 from the World Wide Web: http://www.pbs.org/merrow/tttscript.txt

Miller, Matthew. 1997. Surprise! National school standards exist. *U. S. News and World Report* (November 17). Retrieved December 23, 1997, from the World Wide Web: http://www.usnews.com/usnews/issue/971117/17stan.htm

Milliken v. Bradley. 1974. 418 U.S. 717.

Missouri v. Jenkins. 1995. 515 U.S. 70. Retrieved September 27, 1997, from the World Wide Web: http://supct.law.cornell.edu/supct/html/93-1823.ZS.html

National Center on Educational Outcomes. 1994. *Educational accountability for students with disabilities*. NCEO Policy Directions Number 3. Minneapolis, Minn.: National Center on Educational Outcomes. ERIC Reproduction Document No. ED 378 775. Retrieved November 19, 1997, from the World Wide Web: http://www.coled.umn.edu/NCEO/OnlinePubs/Policy3.html

National Commission on Excellence in Education. 1983. *A nation at risk*. Washington, D.C.: Government Printing Office. Retrieved September 27, 1997, from the World Wide Web: http://www.ed.gov/pubs/NatAtRisk/index.html

Neill, Monty. 1996. Assessment reform at crossroads. *Education Week* (February 21). Retrieved September 27, 1997, from the World Wide Web: http://www.edweek.com

Noble, Audrey J., and Mary Lee Smith. 1994. Old and new beliefs about measurement-driven reform: "Build it and they will come." *Educational Policy* 8 (June): 111-36.

Ohio Association of Independent Schools v. Goff. 1996. 92 F.3d 419 (6th Cir.). Retrieved September 27, 1997 from the World Wide Web: http://www.ljextra.com/cgi-bin/f_cat?prod/ljextra/data/external/1996/08/9608012.c06

Orfield, Gary. 1969. *The reconstruction of southern education: The schools and the 1964 Civil Rights Act*. New York: John Wiley & Sons.

Orfield, Gary. 1993. *The growth of segregation in American schools: Changing patterns of separation and poverty since 1968*. Alexandria, Va.: National School Boards Assocation, Council of Urban Boards of Education.

Orfield, Gary, Susan E. Eaton, and the Harvard Project on School Desegregation. 1996. *Dismantling desegregation: The quiet reversal of Brown v. Board of Education*. New York: New Press.

Peterson, Bob. 1997. We need a new vision of teacher unionism. *Rethinking Schools* 11 (Summer): 1, 3-5.

Peterson, Kenneth D. 1995. *Teacher evaluation: A comprehensive guide to new directions and practices*. Thousand Oaks, Calif.: Corwin Press, Inc.

Ravitch, Diane. 1983. *The troubled crusade: American education, 1945-1980*. New York: Basic Books.

Ravitch, Diane. 1995. *National standards in American education: A citizen's guide.* Washington, D.C.: Brookings.

Reese, William J. 1995. *The origins of the American high school.* New Haven, Conn.: Yale University Press.

Rose, Lowell C., Alec M. Gallup, and Stanley M. Elam. 1997. The 29th annual Phi Delta Kappa/Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan* 79 (September): 41-56.

Roush, Wade. 1996. A census in which all Americans count. *Science* 274: 713-14.

Ruggles, Patricia. 1990. *Drawing the line: Alternative poverty measures and their implications for public policy.* Washington, D.C.: Urban Institute Press.

Sanders model to measure "value added." 1991. *TEA News* 22 (May): 5.

Sanders, William L., and Sandra P. Horn. 1994. The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education* 8: 299-311.

Sapon-Shevin, Mara. 1993. Gifted education and the protection of privilege: Breaking the silence, opening the discourse. In *Beyond silenced voices: Class, race, and gender in United States Schools* (pp. 25-44), edited by Lois Weis and Michelle Fine. Albany, N.Y.: State University of New York Press.

Schell, Jonathan. 1975. *The time of illusion.* New York: Knopf.

Shepard, Lorrie A. 1991. Will national tests improve student learning? *Phi Delta Kappan* 73: 232-38.

Simmons, Warren, and Lauren Resnick. 1993. Assessment as the catalyst of school reform. *Educational Leadership* 50 (February) 11-15.

Sizer, Theodore R. 1992. *Horace's school: Redesigning the American high school.* Boston: Houghton Mifflin Co.

Sizer, Theodore R. 1995. Will national standards and assessments make a difference? In *Debating the future of American education: Do we need national standards and assessments?* (pp. 33-39), edited by Diane Ravitch. Washington, D.C.: Brookings Institution.

Skocpol, Theda. 1991. Targeting within universalism: Politically viable policies to combat poverty in the United States. In *The urban underclass* (pp. 411-436), edited by Christopher Jencks and Paul E. Peterson. Washington, D.C.: Brookings Institution.

Smith, Mary Lee. 1991. Put to the test: The effects of external testing on teachers. *Educational Researcher* 20 (June-July): 8-11.

Smith, Mary Lee, and Claire Rottenberg. 1991. Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice* 10 (Winter): 7-11.

Spring, Joel. 1989. *The sorting machine revisited: National educational policy since 1945* (updated ed.). New York: Longman.

Starr, Paul. 1982. *The social transformation of American medicine*. New York: Basic Books.

Starr, Paul. 1987. The sociology of official statistics. In *The politics of numbers* (pp. 7-57), edited by Paul Starr and William Alonso. New York: Russell Sage Foundation.

Stedman, Larence A. 1996b. Respecting the evidence: The achievement crisis remains real. *Educational Policy Analysis Archives* 4 (April 4). Retrieved September 27, 1997 from the World Wide Web: http://olam.ed.asu.edu/epaa/v4n7.html

Stedman, Lawrence A. 1996a. The achievement crisis is real: A review of *The manufactured crisis*. *Educational Policy Analysis Archives* 4 (January 23). Retrieved September 27, 1997 from the World Wide Web: http://olam.ed.asu.edu/epaa/v4n1.html

Strike, Kenneth A. 1990. The ethics of educational evaluation. In *The new handbook of teacher evaluation* (pp. 356-73), edited by Jason Millman and Linda Darling-Hammond. Newbury Park, Calif.: Sage Publications.

Tennessee Department of Education. 1996. 21st Century Report Card. Retrieved September 27, 1997 from the World Wide Web: http://www.state.tn.us/education/rptcrd96/index.html

Thurlow, Martha L., Judy Elliott, James E. Ysseldyke, and Ron Erickson. 1996. *Questions and answers: Tough questions about accountability systems and students with disabilities*. Minneapolis, Minn.: National Center on Educational Outcomes. ERIC Reproduction Document No. ED 404 802.

Tyack, David B. 1974. *The one best system: A history of American urban education*. Cambridge, Mass.: Harvard University Press.

Tyack, David B., and Elisabeth Hansot. 1982. *Managers of virtue: Public school leadership in America, 1820-1980*. New York: Basic Books.

Tyack, David B., and Larry Cuban. 1995. *Tinkering toward utopia: A century of public school reform*. Cambridge, Mass.: Harvard University Press.

Vinvoskis, Maris. 1988. *An "epidemic" of adolescent pregnancy? Some historical and policy considerations*. New York: Oxford University Press.

Weick, Karl E. 1976. Educational organizations as loosely coupled systems. *Administrative Science Quarterly* 21: 1-19.

Weiss, Carol H. 1988. Interview study. In *Reporting of social science in the national* *media* (pp. 21-171), edited by Carol H. Weiss and Eleanor Singer. New York: Russell Sage Foundation.

Weiss, Janet A., and Judith E. Gruber. 1987. The managed irrelevance of federal education statistics. In *The politics of numbers* (pp. 363-91), edited by Paul Starr and William Alonso. New York: Russell Sage Foundation.

Werum, Regina. 1997. Sectionalism and racial politics: Federal vocational policies and programs in the predesegregation South. *Social Science History* 21: 399-453.

Wiebe, Robert H. 1967. *The search for order, 1877-1920*. New York: Hill and Wang.

Will Washington cut our COLA? 1997. *Solidarity* (January-February). Retrieved October 1, 1997, from the World Wide Web: http://www.uaw.org/solidarity/9701/03b.html

Wilson, Thomas A. 1996. *Reaching for a better standard: English school inspection and the dilemma of accountability for American public schools*. New York: Teachers College Press.

Wirth, Arthur G. 1992. *Education and work for the year 2000: choices we face*. San Francisco: Jossey-Bass.

Wise, Arthur E. 1979. *Legislated learning: The bureaucratization of the American classroom*. Berkeley, Calif.: University of California Press.

Wise, Arthur E. 1994. Choosing between professionalism and amateurism. *Educational Forum* 58: 139-46.

Wise, Arthur E., and Jane Leibbrand. 1993. Accreditation and the creation of a profession of teaching. *Phi Delta Kappan* 75: 133-36, 154-57.

Return to Table of Contents

About the Author

Sherman Dorn

dorn@typhoon.coedu.usf.edu

Sherman Dorn is Assistant Professor in the Department of Psychological and Social Foundations at the University of South Florida. He received his Ph.D. in history at the University of Pennsylvania in 1992 based on his work on the history of dropout policies. He is currently looking at the history of special education in Nashville, Tennessee, from 1940 to 1990.

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Michael W. Apple University of Wisconsin	Greg Camilli Rutgers University
John Covaleskie Northern Michigan University	Andrew Coulson a_coulson@msn.com
Alan Davis University of Colorado, Denver	Sherman Dorn University of South Florida
Mark E. Fetler California Commission on Teacher Credentialing	Richard Garlikov hmwkhelp@scott.net
Thomas F. Green Syracuse University	Alison I. Griffith York University
Arlen Gullickson Western Michigan University	Ernest R. House University of Colorado
Aimee Howley Marshall University	Craig B. Howley Appalachia Educational Laboratory
William Hunter University of Calgary	Richard M. Jaeger University of North Carolina--Greensboro
Daniel Kallós Umeå University	Benjamin Levin University of Manitoba
Thomas Mauhs-Pugh Rocky Mountain College	Dewayne Matthews Western Interstate Commission for Higher Education
William McInerney Purdue University	Mary P. McKeown Arizona Board of Regents
Les McLean University of Toronto	Susan Bobbitt Nolen University of Washington
Anne L. Pemberton apembert@pen.k12.va.us	Hugh G. Petrie SUNY Buffalo
Richard C. Richardson Arizona State University	Anthony G. Rud Jr. Purdue University
Dennis Sayers University of California at Davis	Jay D. Scribner University of Texas at Austin
Michael Scriven scriven@aol.com	Robert E. Stake University of Illinois--UC
Robert Stonehill U.S. Department of Education	Robert T. Stout Arizona State University

---

archives | abstracts | editors | board | submit | comment | subscribe | search

EPAA Commentary

Contributed Commentary on Volume 6 Number 1: Dorn *The Political Legacy of School Accountability Systems*

8 January 1998

Lyle V. Jones
University of North Carolina

lvjones@email.unc.edu

Sherman Dorn presents compelling reasons why we must attend to the political legacy of educational reform when thinking about the pros and cons of national school achievement tests. He emphasizes the ambiguities that surround the targets for accountability: public schools as institutions, teachers and school administrators, students, public policy, etc., many of which cannot be well served by a single index.

Dorn's case is bolstered substantially by a case study of the (failed) efforts in Great Britain, following the passage of the Education Reform Act of 1988, to establish mandatory national tests (see Black, 1994). Recently, I have argued that we are failing to adequately consider the lessons from Britain and from other sources as the nation continues to move towards a program for national testing (Jones, 1997a, 1997b).

Dorn reminds us that an apparently objective notion of a statistical accountability system serves to divert discussion from the purposes of schools and the means by which those purposes may be fulfilled. He correctly concludes that there is high risk that "the domination of crude statistical evaluation of schools will continue, to the detriment of schools, children, their families, and the public."

References

Black, P. J. (1994). Performance assessment and accountability: The experience in England and Wales. *Educational Evaluation & Policy Analysis*, 16, 191-203.

Jones, L. V. (1997a). National tests and education reform: Are they compatible? [On-line]. Available: HTTP://www.ets.org/research/pic/jones.html

Jones, L. V. (1997b). National standards, Yes; national tests, No. [On-line]. Available: http://ericae2.educ.cua.edu//ft//nattest/oped5.htm

EPAA Commentary

Contributed Commentary on
Volume 6 Number 1: Dorn *The Political Legacy of School Accountability Systems*

13 January 1998

Craig B. Howley
Appalachia Educational Laboratory

howleyc@ael.org

Search as one may, one encounters redundant *libraries* of treatises and guides on test uses and even misuses--including uses in policy. I had searched in vain for a critical, contextualized, discussion of high-stakes testing. It seemed merely to be an accepted fact of life. A critical reading of the history and distortions of accountability-as-performance-testing apparently did not exist, until the appearance of this article. Sherman Dorn has done education a good turn with his analyses in "The Political Legacy of School Accountability Systems." Special virtues in this piece include Dorn's clear-headedness about professionalism, sensitivity toward the history and formation of educational discourse, the importance of community context and diversity, and, generally, a steadfast refusal to see in history an inevitable progress.

Dorn's work will be especially helpful in the preparation of an article about accountability in the rural context, which a colleague and I are just beginning. Accountability, it seems, is needed because schools have become so remote from their publics, and the social construction known as the public is itself losing coherence. Rural schools are allegedly very close to their "communities" (their public). Widespread evidence for this claim is much thinner than one would suspect, but in most rural schools, faculty and staff are nearly all local people who interact continually with one another in social and civic encounters outside the school walls. Perhaps this sort of informal phenomenon is what constitutes the oversight for which accountability schemes are intended (unconsciously, of course, in the minds of the framers of such schemes) to substitute. If this is so, the substitution is particularly unsuited to the terrain of rural existence.

Dorn is especially to be thanked, as well, for not demonizing tests. Standardized, norm-referenced tests are both the products, and the poor innocent victims of the technocratic worldview. They are not going away anytime soon, and they can be theoretically helpful in understanding the pattern of a child's accomplishments. Dorn notes the utility of some of these tests for parents of special needs kids; the truth is that most parents could profitably take a similar interest and discover a similar utility. Most teachers of my acquaintance do not, however, find aggregate classroom or school results particularly helpful. They understand the game and they are cynical, widely.

The one usage for which norm-referenced tests, among the gamut of all "standardized" instruments, exhibit wondrous utility is quantitative research. But, of course, bureaucrats, politicians, and government functionaries (a.k.a. "policy makers") have even less respect for researchers than for teachers. More's the pity; but this is a *very useful* article for those with the institutional leisure to write and think about schools.

EPAA • enter the archives • browse the abstracts • the editors • the edit board
• submit article • submit commentary • search • subscribe
volume: • 1 • 2 • 3 • 4 • 5 • 6 • 7

This article has been retrieved 1653 times since January 9, 1998.

Education Policy Analysis Archives

Volume 6 Number 2 **January 9, 1998** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal. Editor:
Gene V Glass Glass@ASU.EDU. College of Education
Arizona State University,Tempe AZ 85287-2411
Copyright 1998, the EDUCATION POLICY ANALYSIS
ARCHIVES.Permission is hereby granted to copy any
article provided that EDUCATION POLICY ANALYSIS
ARCHIVES is credited and copies are not sold.

Review of Stephen Arons's *Short Route to Chaos**

Stephen Arons, (1997) *Short Route to Chaos : Conscience, Community, and the Re-constitution of American Schooling*. Amherst: University of Massachusetts Press 1997, 154 pp plus notes, bibliography, index. ISBN 1-55849-078-7

Charles L. Glenn
Boston University

Stephen Arons, author of *Compelling Belief: The Culture of American Schooling* (Amherst: University of Massachusetts Press, 1986), is one of the most articulate and influential critics of the educational Establishment from the secular Left. In his new book, he takes on the Clinton Administration's efforts to establish national outcome standards--Goals 2000--which he describes as "comprehensive, centralizing, and insensitive to the diversity of goals that students, families, and communities bring to education. Through the use of federal grants and state regulations, it aims to bring every school in every school district in every state into conformity with politically prescribed standards of what should be learned by every child" (page 4). Arons warns that "[o]nce accepted by the public, Goals 2000 will change the balance of power in schoolhouses and courtrooms in a way unlikely ever to be undone. That change in schooling will very likely undermine the freedom of intellect and spirit that has been so essential to the American experience" (page 98).

Over against this threat, Arons sets what he considers the equally menacing efforts of the "Christian Right" to gain control of American schooling in order to undermine freedom. This accusation isn't documented or argued, simply asserted over and over. Is it true that James Dobson (the current bete noir of Progressives) wants to take over the public schools? No, in fact he is calling for vouchers so that parents who wish them can choose religious schools "without financial penalty" as an alternative to public schools. Does Dobson want to reinstitute "school prayer" in the Engel v Vitale sense? Not at all; he recently disavowed that, and wrote that students should be as free to use religious speech as they are to use political or other opinion speech, and no more. The reality is that the "education establishment" which Arons opposes has created the specter of foaming-mouthed ultra conservatives invading the public school, shrine of the American civil religion, to justify its continuing monopoly.

Arons describes a number of recent controversies in which the establishment and the religious Right have struggled over control. Missing from his roster of combatants is the secular Left, which has in fact won far more of the battles to influence the content of the curriculum on issues like sexuality and multiculturalism. It would presumably have been difficult for Arons to admit that the leading cause of resistance by parents to what goes on in public schools has grown out of these victories by the secular Left to shape the message those schools offer. But for Arons, apparently, nothing the Left can do poses a threat to freedom.

Libertarians on the Left, like Arons, are in a difficult position. Most of those who agree with them about the dangers of a government monopoly of education and a strong government role in setting goals for schools are very unwelcome allies: they are conservative Christians whose views they find highly distasteful.

> Among the most frequent targets are secular humanism, the separation of church and state, Darwinian evolution, sexuality and health education. There is little tolerance for any worldview other than that of heterosexual, white, middle-class Christians of Western European origin; little respect for freedom of expression among students and in student publications; and in general, antagonism toward teachers and students who try to explore and evaluate life's most challenging problems of personal, social, or moral conduct. (page 55)

On the other hand, Arons also wants to distance himself from the critics of religious conservatives, as when he points out that People For the American Way's report on censorship efforts "did not even mention that the original selection of textbooks--by statewide, politically created government agencies in twenty-three of fifty states, for example--is as much an act of censorship as the effort to remove those materials once they have been selected" (page 57).

So whom does Arons like and admire? Groups of parents and others who hold contrarian views about how they want their children educated, like the Satmar Hasidim in the Kiryas Joel case in New York State, who can be romanticized because they are exotic and do not relate to anything that can be perceived as threatening potentialities in American life. But not conservative Catholics and Protestants, the people who supported Pat Robertson. Unfortunately for his proposal to "re-constitute American schooling" on the basis of community and the free-exercise of conscience, it is obvious that the great majority of new schools that would spring up under a free and equitable system of educational funding would be based on religious convictions that most Progressives would find very distasteful indeed. That's what freedom's about.

Arons's opposition to centralization does not lead him to support a return to more local control of schools, which he sees as equally unfavorable to freedom: "like Goals 2000, local control can secure neither freedom of intellect and belief nor equal educational opportunity in public schools. It can advance neither the empowerment of parents and communities nor the professionalism of teachers. It can neither reduce unnecessary conflict over matters of conscience nor increase the overall quality of education available to American children" (page 103).

So what does Arons want? He has four concrete and sensible proposals: school choice, school and teacher independence from government regulation of instructional content, a right to publicly-funded schooling, and equity in funding (page 144). These proposals deserve to be spelled out, and the appropriate cautions (consumer protection, for example, and equal access) and nuances inserted. It would have been helpful if Arons--a legal scholar--had confronted the difficult legal issues that would arise under a system of real educational freedom. For example, should schools be entitled to discriminate on the basis of religion, philosophy, sex, or race in admitting pupils? In hiring staff? In dismissing staff who exercise their "academic freedom" in ways contrary to the distinctive character of the school? If not, how can schools preserve this distinctive character? And if they cannot, will real choice exist for parents who want schools with such a character, and for teachers who want to teach in such schools? What about the pupil who questions received authority, in a school which has been chosen by parents and teachers who want education based upon such authority?

Arons devotes almost no effort to justifying his proposals or to showing how they might be worked out, but turns immediately to calling for a national discussion that would, he believes, lead us to a new level of understanding and an education amendment to the U.S. Constitution. Although conceding that this might "seem an unfavorable time because the Education Empire and the Christian Right continue to be locked in battle over ideology, power, and self-interest in the schools" (page 148), Arons insists that "ordinary citizens" can and must "seize the constitutional moment and depoliticize public education" (page 145). It is exceedingly hard to see how such a discussion could--or should--take place in a democratic society without being "political," nor does Arons offer any suggestions about how it might take place, or under whose sponsorship. A sort of communitarian fuzziness afflicts this erstwhile Libertarian.

Short Route to Chaos is unfortunately not an especially convincing case for the dangers of government control of education through national standards. That such a case could be made, there can be no doubt, but it would have to show how such standards would enforce more conformity than already exists as a result of professional norms and the economics of textbook publishing. In fact, comparative studies have found that schools in France and other countries with national standards enjoy more real autonomy than do schools in the United States, subject as they are to oversight and interference by more than fifteen thousand local school boards. Of course, in France and most other democracies parents can choose publicly-funded non- government schools for their children, including religious schools. This support for freely-chosen community--for which Arons makes an eloquent case--does not appear to conflict with the national education standards which most of these countries have also adopted.

Americans are re-assessing a system of schooling which makes less provision for conscience and community than do those of other countries. Most of the impulse for this reassessment comes from the disenchantment of parents with the quality and with the prevailing secularism--rather than religious neutrality--of public schools. Stephen Arons brings an important contrasting perspective which reaches the same conclusions from a very different starting point. It seems likely, however, that it will continue to be through *Compelling Belief* rather than *Short Route to Chaos* that his voice will be heard.

* Stephen Arons responds to this review in the next article.

About the Author

Charles L. Glenn

Email: glennsed@bu.edu

Phone: 617/353-7108

Charles L. Glenn (AB, EdD, Harvard University; PhD, Boston University) is the Chairman, Department of Administration, Training, and Policy Studies and Professor of Education in the Boston University School of Education.

The formulation and implementation of policies affecting the education of urban and racial, ethnic, and religious minority students are the focus of Dr. Glenn's teaching and research. He has published extensively on parent choice, desegregation, use of minority languages in schools, and religion and education. *The Myth of the Common School* (1988) is a historical study of resistance to efforts to use schooling to reshape society in France, the Netherlands, and the United States. *Choice of Schools in Six Nations* and *Educational Freedom in Eastern Europe* survey current policies and controversies. His forthcoming *Minority Languages in Schools* considers how twelve industrialized nations educate the children of immigrants. Dr. Glenn directed the Massachusetts Department of Education's equity and urban-education efforts for more than 20 years and continues to work with educational systems in the United States, Eastern and Western Europe, and the Middle East on policies to balance common standards with school-level autonomy and choice by parents and teachers.

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

Greg Camilli
Rutgers University

John Covaleskie
Northern Michigan University

Andrew Coulson
a_coulson@msn.com

Alan Davis
University of Colorado, Denver

Sherman Dorn
University of South Florida

Mark E. Fetler
California Commission on Teacher Credentialing

Richard Garlikov
hmwkhelp@scott.net

Thomas F. Green
Syracuse University

Alison I. Griffith
York University

Arlen Gullickson
Western Michigan University

Ernest R. House
University of Colorado

Aimee Howley
Marshall University

Craig B. Howley
Appalachia Educational Laboratory

William Hunter
University of Calgary

Richard M. Jaeger
University of North Carolina--Greensboro

Daniel Kallós
Umeå University

Benjamin Levin
University of Manitoba

Thomas Mauhs-Pugh
Rocky Mountain College

Dewayne Matthews
Western Interstate Commission for Higher Education

William McInerney
Purdue University

Mary P. McKeown
Arizona Board of Regents

Les McLean
University of Toronto

Susan Bobbitt Nolen
University of Washington

Anne L. Pemberton
apembert@pen.k12.va.us

Hugh G. Petrie
SUNY Buffalo

Richard C. Richardson
Arizona State University

Anthony G. Rud Jr.
Purdue University

Dennis Sayers
University of California at Davis

Jay D. Scribner
University of Texas at Austin

Michael Scriven
scriven@aol.com

Robert E. Stake
University of Illinois--UC

Robert Stonehill
U.S. Department of Education

Robert T. Stout
Arizona State University

---

archives | abstracts | editors | board | submit | comment | subscribe | search

This article has been retrieved 1053 times since January 9, 1998.

Education Policy Analysis Archives

Volume 6 Number 3 January 9, 1998 ISSN 1068-2341

A peer-reviewed scholarly electronic journal. Editor:
Gene V Glass Glass@ASU.EDU. College of Education
Arizona State University,Tempe AZ 85287-2411

Copyright 1998, the EDUCATION POLICY ANALYSIS
ARCHIVES.Permission is hereby granted to copy any
article provided that EDUCATION POLICY ANALYSIS
ARCHIVES is credited and copies are not sold.

Planting Land Mines In Common Ground:
A review of Charles Glenn's review of *Short Route To Chaos*

Stephen Arons
University of Massachusetts at Amherst

Abstract Arons responds to what he considers to be Glenn's misrepresentations of the tone and content of *Short Route To Chaos*. He writes that Glenn "appears to be attempting to construct the book's message into just one more salvo fired in the endless school wars. It is anything but....Reading Glenn's review, one is left with the impression that the book is a Christian-bashing, left-leaning, work of communitarian fuzziness in which a legal scholar unaccountably refuses to confine himself to ... technical explication of existing constitutional doctrine." In his response, Arons affirmatively sets out some of the book's main themes of political /cultural conflict over standardized schooling, corrects some of what he sees as Glenn's misundertsnadings, and notes that the book itself invites readers to eschew partisanship and recognize that there are deep structural problems in American public education. In closing, Arons uses an example of Glenn's partisan misunderstanding that leads Arons to recommend to the reader that it would be better to read *Short Route to Chaos* for oneself.

One of the central themes in *Short Route To Chaos* suggests that unless Americans step back and get some perspective on the current school wars and the endless rounds of school reform fads, we are destined to keep repeating the nearly neurotic cycle of conflict that has characterized public schooling since the mid-nineteenth century. This conflict is unnecessary, culturally corrosive, and increasingly destructive of school quality; and the book discusses frankly the ugly side of these battles. Having done so, *Short Route To Chaos* invites the reader to put aside partisanship long enough to see that there are deep structural problems in American public education which are themselves a primary cause of this perennial conflict over the content of schooling.

Early in the book, while stories of the school wars are still being told and the analysis of their causes has not yet become the focus of the work, I try briefly to deal with the problem of holding a mirror up to the unseemly spectacle of the American preoccupation with school wars. Noting that the Christian Right has been very successful at exploiting the weaknesses in public education structure, the book states:

> But the Christian Right meets its match in an education establishment artful at demonizing its opponents and willing to resist virtually any attempt to change the ideology and practices of the public schooling to which it owes its existence...
>
> So effectively do these two giants demonize each other, and so distorted has the public debate over schooling become as a result, that it is difficult to discuss the attack by the Christian Right or the defensiveness of the education establishment without seeming to insult large numbers of well-intentioned people on both sides. 'Right-wing Christian' is to most Christian fundamentalists, for example, as 'tax-and-spend liberal' is to many other Americans of good will: a label, a stereotype, a mischaracterization of citizens trying to improve the quality and meaningfulness of public schooling for their children and their community.
>
> It is essential to get beyond the demonization and polarization, and to put in perspective the partisan attacks on public schooling and the hackneyed defense of the status quo there. Americans with conflicting but sincere views about schooling need to admit that some leaders on each side have been willing to misuse the legitimate concerns of their constituents....

But it isn't easy to escape the lure of immediate self-interest and ideological commitment in these conflicts, as Professor Glenn's accompanying review of *Short Route To Chaos* makes clear. When what is at stake is so important, and when both the school wars themselves and their spoils seem to provide so much satisfaction, even the most astute may find it difficult to see beyond the end of their own agendas. If the book is an invitation to cease politicizing American education for a moment and to look squarely at the structural problems of public education, then Glenn has either not understood this invitation or has intentionally chosen not to take it up. This is an unfortunate posture for a scholar, though it is completely understandable in a partisan.

The U.S. Supreme Court warned of the problem of politicized schooling and the chaos of conflict long ago in West Virginia v. Barnette:

> As governmental pressure toward unity becomes greater, so strife becomes more bitter as to whose unity it shall be. Probably no division of our people could proceed from any provocation than from finding it necessary to choose what doctrine and whose program public educational officials shall compel youth to unite in embracing.

More than fifty years later, as the state and federal governments begin trying to standardize American education along the lines of the Goals 2000: Educate America Act, the bitter, predictable strife continues and increases. Among its chief casualties has been freedom of conscience in education--the individual liberty to follow an internal moral compass in setting a course for a meaningful and fulfilling life. Undermined as well has been the building of community, which most teachers and families believe to be essential to successful schooling. Hence the subtitle of Short *Route To Chaos: Conscience, Community, and the Re-Constitution of American Schooling.*

I argue in *Short Route To Chaos* that schooling has become so burdened with unnecessary conflict that it is becoming increasingly dysfunctional. It is therefore in all our interests--not just the Christian Right or the secular left--to reduce the level of political/cultural warfare over schooling. This is analogous to reducing the level of conflict over religion 200 years ago when the Bill of Rights adopted the requirement of separation of church and state.

Another important theme of *Short Route* is the importance of focusing the public debate on the principles by which public education should be organized, rather than on the specific programs or proposals advanced by one partisan group or another. That is why the book stresses the "constitutional" level of reform, suggesting an extended national dialogue on an education amendment to the U.S. Constitution. Under the present conditions, to debate vouchers or charter schools or decentralization or home schooling would be to fall back into the old cycle of partisan conflict. But once we agree on fundamental principles--or at least start discussing them civilly--the appropriate mechanisms for achieving them are likely to become clear and to come within our reach.

Glenn ignores this theme as well, apparently preferring to criticize *Short Route* for not providing a detailed defense for what he imagines I would advocate as a suitable program, vouchers. Given Glenn's past thoughtfulness about matters of education policy, it would have been more useful for him to have joined instead in a discourse about the basic principles underlying vouchers, rather than the programmatic details of this or any other program that might eventually be advanced. Here is what the last chapter of the book says about the difference between principles and programs:

> ...a constitutional amendment for education cannot spell out a particular program for schools. It must, like the Bill of Rights, be based upon a few principles which specify government

> powers, secure fundamental freedoms, and establish the ground rules under which particular programs may be created, put into service, and judged for their constitutionality. (p.149)

Perhaps professor Glenn took umbrage at the book's introductory comment that "It is...my intention to suggest how the American people themselves--not limited by the current views of their political representatives, education experts, and constitutional courts, and quite apart from 'politics as usual'--may achieve a re-constitution of schooling adequate to strengthen both conscience and community in public education." (P.10)

Glenn's misunderstandings and misrepresentations extend still further. He claims that *Short Route* does not make "an especially convincing case for the dangers of government control of education through national standards." Glenn can be forgiven for not agreeing with Theodore Sizer's estimate that "Arons' argument is politically very incorrect, but devastating." But it appears that he has not read the chapter in which the argument he dismisses is centered, "Renouncing Our Constitutional Heritage." More interesting, however, is what Glenn would consider to be the basis of a convincing argument: "it would have to show how such standards would enforce more conformity than already exists as a result of professional norms and the economics of textbook publishing."By this standard, I suppose that a theocratic state would be acceptable as long as the majority of its citizens shared the religious beliefs of their rulers.

The primary danger of government control of school content--through politically-defined education standards, testing programs, or other means--is not conformity. The danger is that in giving government at any level the power to control school content, we invite endless and destructive political conflict over whose idea of good education will be adopted by the state. That, in effect, is what the Supreme Court meant when it declared, in the Barnette case, that the "ultimate futility of...attempts to compel coherence is the lesson of every such effort...Compulsory unification of opinion achieves only the unanimity of the graveyard." Moreover, in empowering government to control school content we risk renouncing a constitutional heritage which holds that, in matters of intellect and belief, government has no proper role beyond protecting individual liberties. That, to quote professor Glenn, is "what freedom's about."

There are other themes in *Short Route To Chaos* that Glenn either ignores, misunderstands, or misstates--that American public schooling has already been re-constituted by state and federal laws adopted without meaningful public debate; that conscience and community are not mutually exclusive but mutually dependent; that schooling is much more like religion than it is like economic policy or public policy; that the Christian Right and the Education Empire are equally destructive and unattractive in their campaigns to get or hold power over schooling; and that the Education Empire--including Glenn--is more a part of the problem than of the solution.

Reading Glenn's review, one is left with the impression that the book is a Christian-bashing, left-leaning, work of communitarian fuzziness in which a legal scholar unaccountably refuses to confine himself to the kind of technical explication of existing constitutional doctrine that a conservative Christian could use for partisan purposes in the school wars. I don't mind controversy; and argument is my stock in trade. But knowing Charles Glenn's past commitment to freedom of conscience and to equality of educational opportunity, I expected a more thoughtful dissent. In closing, therefore, I offer one example of partisan misstatement that particularly galled me and that, I hope, illustrates why it would be better to read *Short Route To Chaos* for oneself than to be satisfied with Glenn's dismissive and combative review.

Glenn criticizes the book for trashing the Christian Right but admiring the Satmar Hasidim of New York "who can be romanticized because they are exotic and do not relate to anything that can be perceived as threatening potentialities in American life. But not conservative Catholics and Protestants,..." Here is what *Short Route To Chaos* actually says about the Satmar and the Kiryas Joel case:

> The Court simply could not accommodate the legitimate claims of the Satmar and simultaneously uphold the principles of the Establishment Clause. But had it been parents instead of governments that chose where each child attends an approved school, the Court's dilemma would have dissolved.
>
> Without such a structural change in schooling, however, any accommodation acceptable to the Satmar and approved by the Court would have been so narrowly drawn that it would likely be virtually useless to other communities--including many Christian fundamentalists, who are no less entitled to respect for their community and religious values than the Satmar [emphasis added]. The lesson of this long struggle therefore seems clear. Public schools are presently structured so that they become the enemies of private conscience and the building of communities of belief. Making it easier for schooling to be consistent with any community's most basic beliefs is a problem that can be solved by restructuring public education, not by reinterpreting the First Amendment.

Whether by design or by inadvertence, Charles Glenn has misrepresented the tone as well as the substantive themes of *Short Route To Chaos*. In so doing he appears to be attempting to construct the book's message into just one more salvo fired in the endless school wars. It is anything but. The school wars are ugly; and they do bring out the worst in many well-intentioned Americans, as they undermine the quality of schooling and the vitality of both conscience and community. If we are ever to ameliorate this destructive conflict, we must have a truce just long enough to see how needlessly we are pitted against each other by a school structure that simultaneously apportions freedom of choice according to wealth and requires majority consent for the exercise of individual conscience.

There are pragmatic and principled solutions available if we can just stop planting land mines in what could be our common ground. Perhaps

Charles Glenn would rather fight than solve problems. But that approach will get us nothing more than another 150 years of school wars. But if he so chooses, Glenn has the ability and the experience needed to help call a truce and to find solutions that respect diversity. Perhaps he still will.

About the Author

Stephen Arons

Department of Legal Studies
University of Massachusetts at Amherst

Email: arons@legal.umass.edu

Phone: (413) 545-3536

Stephen Arons, B.A. Univ. of Pa., J.D. Harvard, is professor of Legal Studies at the University of Massachusetts, Amherst, and a member of the Massachusetts Bar. For the past 25 years, Arons has been involved in issues of schooling, public policy, and constitutional law from a number of different perspectives. He was one of the founders of an alternative school for street youth, worked as a staff attorney concerned with civil rights at the Center for Law and Education, was an early participant in the federal study of school vouchers, wrote extensively about education policy for the *Saturday Review* and other magazines, has litigated issues ranging from state aid for private schooling to parental rights in home education, and has consulted for state and federal departments of education and legislative committees concerned with the constitutional implications of various school finance mechanisms. Arons has written numerous articles in professional journals and two books on schooling, culture, and the U.S. Constitution: *Compelling Belief: The Culture of American Schooling* (1986) and *Short Route To Chaos: Conscience, Community, and the Re-Constitution of American Schooling* (Univ. of Mass. Press, 1997).

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

Greg Camilli
Rutgers University

John Covaleskie
Northern Michigan University

Andrew Coulson
a_coulson@msn.com

Alan Davis
University of Colorado, Denver

Sherman Dorn
University of South Florida

Mark E. Fetler
California Commission on Teacher Credentialing

Richard Garlikov
hmwkhelp@scott.net

Thomas F. Green
Syracuse University

Alison I. Griffith
York University

Arlen Gullickson
Western Michigan University

Ernest R. House
University of Colorado

Aimee Howley
Marshall University

Craig B. Howley
Appalachia Educational Laboratory

William Hunter
University of Calgary

Richard M. Jaeger
University of North Carolina--Greensboro

Daniel Kallós
Umeå University

Benjamin Levin
University of Manitoba

Thomas Mauhs-Pugh
Rocky Mountain College

Dewayne Matthews
Western Interstate Commission for Higher Education

William McInerney
Purdue University

Mary P. McKeown
Arizona Board of Regents

Les McLean
University of Toronto

Susan Bobbitt Nolen
University of Washington

Anne L. Pemberton
apembert@pen.k12.va.us

Hugh G. Petrie
SUNY Buffalo

Richard C. Richardson
Arizona State University

Anthony G. Rud Jr.
Purdue University

Dennis Sayers
University of California at Davis

Jay D. Scribner
University of Texas at Austin

Michael Scriven
scriven@aol.com

Robert E. Stake
University of Illinois--UC

Robert Stonehill
U.S. Department of Education

Robert T. Stout
Arizona State University

---

archives | abstracts | editors | board | submit | comment | subscribe | search

This article has been retrieved 940 times since January 30, 1998.

Education Policy Analysis Archives

Volume 6 Number 4 January 30, 1998 ISSN 1068-2341

A peer-reviewed scholarly electronic journal. Editor:
Gene V Glass Glass@ASU.EDU. College of Education
Arizona State University,Tempe AZ 85287-2411
Copyright 1998, the EDUCATION POLICY ANALYSIS
ARCHIVES.Permission is hereby granted to copy any
article provided that EDUCATION POLICY ANALYSIS
ARCHIVES is credited and copies are not sold.

Comparative Issues of Selection in Europe: The Case of Greece

Dionysios Gouvias
University of Manchester

Abstract

This article deals with inequality of access to higher education in Greece, and especially, the case of the Metropolitan Area of Athens. Specifically, I deal with a general overview of the debates about "selection" in the educational systems of Europe, with special reference to the case of Greece. It is argued here that in those levels of the educational "ladder" where the degree of specialisation and the need for individual selection is insignificant, inequalities exist, but are not profound. On the contrary, in the upper levels, and especially as the time to enter (or to be trained to enter) the labour market comes closer, students' success depends much more on externally assessed examination performance and, therefore, a more rigorous selection process emerges--a process that is decisively influenced by the labour- market requirements and limitations. Finally, an extended examination of the evolution of the Greek school-system and the changes in examination practices, and the relationship between the structure of the school system and the job-market, will be attempted.

General Framework

During the 1950s and 1960s, the study of education-- until then dominated by the traditional "individualistic" values of "excellence" and "merit"--became more closely associated with the social scientific approach. That is not surprising, if one considers the context of education on a global scale, after the Second World War. All disciplines included in the so-called social sciences domain (especially sociology and psychology, and very often economics as well) faced highly controversial problems concerning the consequences of the rapid growth in school enrolment rates that characterized most countries. The enrolment explosion at the secondary school level and expanded admissions to university preparatory schools as well as to the university itself have given rise to questions about the "quality" of students processed through a system of mass education, as compared to an elitist one. In a selective system, children are allocated to different types of school at early ages by means of organisational differentiation. Also at an early stage in their school careers, grouping practices are employed to spot those who are supposed to be particularly academically oriented.

Nevertheless, selection is not--as has been argued in the past--only about "sorting out" the "ablest" or "academically oriented" pupils. At the same time it is a general social phenomenon, an indispensable aspect of the existence of human societies. From the employment of a job-candidate, to the election of a party- leader, selection procedures are always followed to arrive at a final choice. In that sense, selection is not only unavoidable, but also crucial for every social function. Sociologists since Durkheim's age have concluded that school contributes to the continuity of "social balance", by transmitting to the new generations rules, principles and moral categories, which reflect the society's "views" about what is good and bad, progressive and conservative, and the like. The controversy starts when one questions the legitimacy of those values and principles as serving specific interests of specific social "groups", or "classes", or "layers" of society. In an ideal society, only the inherited abilities of an individual would define his or her position within the social system, and that could start from the very early stages of socialization, including socialization to school. However, many factors other than the ability of the students influence their eventual educational experiences and attainments. These includes differences in the level and quality of education available in the country, region, or community in which they live; differential access to educational facilities according to their social class, religion, race and ethnic origins; differences in the willingness and abilities of their parents and others to provide the financial and psychological support necessary for the maximization of their potential talents.

Theoretical Debate on the Relation Between Asessment and Selection

As Wood (in Gipps and Murphy, 1994, p.40) suggested, the definition of equal opportunities includes: 1) equal life chance, 2) open competition for scarce opportunities, 3) equal cultivation of different capacities and 4) independence of educational attainment from social origins.

However, one could hardly ever argue that all the above aspects of the notion of "equal opportunity" were consistently taken under consideration in the policy- planning process of various European educational systems in the past. In each European country, there have developed different types of examination practices, at national, regional and local levels, based on school or on external assessment, depending on the structural elements of each system, and on the relationship between education and society as a whole.

The move through the centuries, from the monastic discipline of the Middle Ages, to the "refined" and "noble" ideas of "humanism" and "cultivated spirit" during the Enlightenment, and then to the intensely competitive system of hierarchically organized instruction (initiated by the Jesuits in order to adapt their religious purposes to the progressively individualistic social climate of the Industrial Revolution), followed patterns already evident in global societal changes. As Durkheim argued "it is no accident that competition becomes more lively and plays a more substantial role in society as the movement towards individualisation becomes more advanced." (Durkheim [1969], in Karabel & Halsey, 1977, p.105)

Indeed, the "industrialisation" of the western Europe, and the enormous expansion of trade on an international scale, influenced the educational systems by "injecting" into them a more "utilitarian" set of values, and by making them more open to competition, which it was thought would enable the "ablest" to prevail.

The unprecedented changes brought on by the Industrial Revolution had enormous economic, political and broader social effects. The great mass of people, rising from modest, though rarely very deep, poverty, and the even greater mass of those pressing below them out of the labouring poor into the middle classes, were too numerous to be absorbed. They came to think of themselves increasingly as a "middle class", and not merely as a "middle rank" in society. They claimed rights and power; subsequently, they sought better education for their children. (See Hobsbawm, 1968, pp.79-96) Moreover, on clearly ideological grounds, liberal from the 18th century (like Adam Smith) had been calling for reform in favor of the rising middle classes, and for selection by "merit". Thus, the "fairest" way to select students seemed to be by the introduction of a widespread system of examination and certification, which would be monitored and controlled by various experts bodies at the national or local level.

In other more centralized educational systems, such as those of Russia and France, any kind of school selection was from very early on directly influenced and controlled by the State mechanism. In the former, the move from the Tsarist "impenetrable polity" to the Bolshevic "socialist" regime after the 1917 Revolution, ensured that political manipulation would remain the principal source of change, no matter how great the differences in social philosophy and political goals. (Archer, 1979, pp.284-306) In the latter, the existence of a highly centralized bureaucracy from the time of Napoleon, not only offered the most prestigious professional opportunities, but also affected enormously the "selection" practices, under a nationally homogenous system of organization, supervision and certification. Examination practices--as indeed most of the school practices--were "dictated by the Baccalaureat, circumscribed by the standardised curriculum and supervised by the Conseil de l' Universite and the inspectorate". (ibid., p.307)

Systematic criticism of the "distortive" distinction between success and failure that the examinations produce, started very early in Europe, and it reflected functional, methodological and sociological concerns. The target of that criticism was mainly the "selection" aspect of examinations, and its "side-effects" on the curriculum, the learning processes (i.e. promotion of uncritical memorization) and the psychological development of each individual pupil (i.e. stress and confusion resulting from a strong competitive environment).

As far as the curriculum was concerned, despite the widely accepted principle that the examination content should reflect the curriculum content, there have been numerous examples of the reversed happening in the past. The existence of "subject groups" or "branches of study", of the upper-secondary school in most of the European systems, reveals the "dependence" of the curriculum on the examination requirements. (Polydorides, 1990, p.87) Instead of having examinations assessing the (achievement in a given) curriculum, what happens is that the curriculum is "adopted" to the specific requirements and limits that a certain assessment system imposes, usually as mandatory rules or guidelines.

In a few words, the advantages and disadvantages of the formal assessment systems (school-based or "external"), as they have been identified by the research community (see Broadfoot, 1979; Wood, 1987; Gipps and Murphy, 1994) can be summarised as below:

Advantages

- elimination of the influence of "luck".
- adoption of different assessment procedures for different student "potentials".
- homogeneity of practice, since the assessment is made according to common criteria.
- effective administration of procedures.
- smaller danger of confusion in school level.

Disadvantages

- lack of account of the internal school practices.
- very "narrow" perception of the notion of "adequate school achievement".
- danger of prejudice against certain social and ethnic sub-groups.
- limited "descriptive", and mainly "interpretive" results, and therefore inadequate information for remediation or improvment.

In addition, there have been strong arguments against the more "extreme" face of assessment: the standardized tests. As Wood points out, "the notion of the standard tests a way of offering impartial assessments of course a powerful one, though if there is not equality of educational opportunity preceding the test, then the "fairness" of this approach is called into question" (quoted in Gipps and Murphy, 1994, p.15) One of the concerns expressed frequently by various researchers is that we are unlikely to know that we have provided equal opportunities until we get equal outcomes. But, if equal opportunities relates to not putting obstacles in the way of particular groups, it does not follow necessarily that factors such as interest, diligence, relevant experience, socioeconomic, cultural and linguistic environment will be equal among groups. In other words - and despite a lack of consensus - there seems to be a general understanding that formal "equality of opportunity" is not sufficient to ensure fairness, nor that striving for "equality of outcomes" is sound, since "different groups may indeed have different qualities and abilities and certainly experiences." (Gipps and Murphy, 1994, p.17)

It is these issues that often raise the problem of "bias" and "validity". "Bias" is generally taken to mean that "the assessment is unfair to one particular group or another".(ibid., p.18) Of course this very general definition does not necessarily address the construction of the various standardised tests, but rather it stresses the unsuitability of some tests for specific measurements. In other words, differential performance on a test by different social groups may not be the result of bias in assessment; it might have been caused by real differences in performance among groups, which may in turn be due to differing access to learning or differing life experiences. If one accepts that differences in interest and motivation are considered to be biasing factors, all tests or assessment methods may be said to have a certain amount of bias. Often in the past, American policy makers, under the pressure of "affirmative action" in the last three decades, tried to manipulate test items and devise tests which favored blacks over whites. Certain kinds of problem are being encountered recently by English counterparts, in the latter's attempts to deal with the increasingly controversial issue of "adequately" - and at the same time "fairly" - assessing the performance of ethnic minorities.

"Validity" is closely related to "bias," although it has a more "technical" connotation. It is generally seen as the extent to which an assessment tool - often a standardised test - measures what it claims to measure. In that sense, one can easily have a test which, according to certain criteria ("criterion validation"), may be claimed to be "valid", but at the same time might be claimed inappropriate, or irrelevant, or meaningless, to a certain sub-group of test-takers (lack of "content validity").

Today, in the midst of an international trend towards standardisation of assessment procedures (see for example the research carried out by the International Association for the Evaluation of Educational Achievement), one overriding fact must not be forgotten: differences in group performance may be due more to the environmental, psycho-social influences that impinge on groups of pupils, or considerably affect the content, administration and scoring of tests, than to any sort of hereditary ability.

School Structures in Europe

Europe today is in the midst of a process of socioeconomic unification in its western regions and a desperate struggle for national identification (as an effect of the late 1980s disintegration) throughout its eastern and south- eastern parts . Despite the fact that one cannot possibly speak of a unified entity under the name "Europe," today), it is of great importance for one to see how selection mechanisms operate in this geographic whole. From the following brief account, I hope that many useful comparisons may be derived between the context in which the Greek educational system developed and currently operates. It is hoped that these comparisons help clarify various elements of the Greek system and serve as a guide for future analyses of it

Unavoidably, we must focus on Western Europe, due to the availability of data and the possible familiarity with some of the relevant educational systems. Moreover, the guiding principles of Western European systems affected to a great extent the decision-making process and the orientation of the Greek educational policies in the last 100 years.

Reform movements in Western European education have gradually gathered momentum throughout the region (Note 1) since the stock-taking days of the immediate post-war period. All countries without exception found themselves faced with the same problems. And these problems all turn on the single fact that the number of children seeking some form of post-primary education has grown out of all proportion to the birth rate during the post-war years. As technology has developmented, the need for greater social mobility has been recognised. In addition to a "basic" schooling, linked to primary education provision, it has been commonly held that there must be an "undifferentiated secondary education, with an integrated curriculum to replace the former crazy patchwork of differentiation" (Mallinson, 1980, p.67).

Traditional secondary academic schools might be able to withstand change so long as students were recruited from the same "upper-middle" and "bourgeois" class, but even then as student protests of the late 1960s demonstrated-- the curriculum was also compelled in some measure to conform.

There have been many commonalities in how these systems evolved to the present day. For example, fees in publicly maintained secondary schools - at least for the compulsory part - were abolished, and many independent institutions came to arrangements whereby they also in certain circumstances could provide free secondary education. All post-primary schools which still had continued to function as a reminder of the old "dual system" ("dual" in technical, as well as in social terms) were upgraded to the secondary level. Flexible arrangements-- under national, regional or local initiatives--based on the principle of popular "enlightenment" through liberal studies, have been made in adult education, and measures have been taken for the examination and certification of people who had never before had formal schooling. In response to pressures from industrial and commercial organizations for improved links between the formal educational instruction and the requirements of a demanding working environment, a number of apprenticeship schemes (in-work training, with part-time attendance of a vocational course in school) have been set up. Lastly, examination hurdles formerly placed at the completion of a child's primary school course to decide to what type of secondary education she or he should attend, were swept away. Everybody was to have the right to some kind of secondary education, at least up to the end of compulsory schooling.

Of course, differences between the various systems never ceased to exist in certain key characteristics: the multitude of alternative "paths" after basic schooling, definitions of what constitutes "primary" and "secondary" levels, starting and leaving age for compulsory schooling, possible charging of fees at a certain level of schooling, opportunities for apprenticeships, degree of centralization of control on administration or curriculum policy, and the like. These issues had to be dealt with in each country individually and as quickly as possible since two major problems arose:

- The post-war baby-boom in combination with the aforementioned influences caused pressing demands for new buildings, teaching materials and enough teachers to deal with the varying needs of the new influx. While in the past, children who had not properly mastered certain "basic skills" by the end of primary school either never sought secondary education or were held back until the skills had been mastered, it was now thought that the "mass" secondary school had to be remedial and make up deficiencies in such skills before any secondary course could be of any worth.

- A very large number of children would abandon school as soon as they reached the compulsory age limit. Therefore, attention should be paid to preparing those children for the world of work.

How the above issues have been dealt with by the various national school systems and how the selection mechanisms have been modified to satisfy the post-war pressures for better schooling (qualitatively and quantitatively) and equal opportunities, will be developed below.

1. The Scandinavian System

The model of schooling in the Scandinavian countries (including Denmark) is characterised by the promotion and, to an impressive extent, implementation of the "comprehensive" ideal. This ideal entails a dynamic approach to the fast- growing needs created by the "triple explosion" of population, of knowledge and of aspirations, which in turn were the results of an "exceptionally rapid urbanization which all too soon revealed how under- developed and uneducated" was the populus during the 1920s and 1930s. (Mallinson, 1980, p.173). Focusing on the Swedish and Danish systems, we may Construct a prototype of the "Scandinavian" school system--although one must be careful not to overgeneralize these observations.

One of the paths created in the Scaninavian system is That in which primary education is linked to lower-secondary in forming a somewhat "unified" and extended nine-year "elementary" school (folkeskole in Denmark). The role of this school is the integration of basic schooling (teaching of "arithmetic/mathematics", native and in the later stages one foreign language, religious education, familiarization with modern literature) with an introduction to vocational studies. (see Elvin, 1981, pp.48-52)

The upper-secondary school, usually starting at 16, is then divided into a general education section (gymnasium) and a vocational education and training path. The former traditionally prepares students for higher education, and the latter qualifies them for work in trade and industry.

Another aspect of this model is that the gymnasium is not the only path to higher education. In Denmark, there is another type of preparatory course for entry to higher education, the "Higher Preparatory Examination" course, which takes two years (the former takes three) and entitles anyone who attended it - even in a county adult center - to participate in the relevant examination. Thus, apart from the traditional way of gaining access to higher education, these reforms (the HF was established in 1966) permitted more mature students "who have already experienced the employment market" to share the opportunity for tertiary education. (Winther-Jensen, in Brock & Tulasiewicz, 1994, p.53)

In Sweden there is no longer a school-leaving examination or test for entry to some form of higher education, since these were all replaced by a certificate (slutbetyg) which lists the average mark per subject (out of a maximum of 5) attained by the student. As Mallinson informs us, in the early 1980s this reform was followed by an increase in the number of graduates of gymnasium entering some further study (Mallinson, 1980, p.177).

2. The "Benelux" Countries

The structure of the educational systems of these countries is characterised by an influx of interconnected and balanced pathways towards either academic higher education or adequate preparation for working life. The distinction between general and technical or vocational education is rather blurred, since not only the different sub-types of secondary school offer a number of "specialisations" from a very early stage (immediately after the completion of primary school), but also the existence of a kind of "transition" period that enables the administration--as well as the pupils and their families--to choose the "best" way forward.

In Belgium, after the reforms in 1969, secondary education (6 years) was divided into 3 cycles, each of two years' duration. The first cycle constitutes a period of observation, the second a period of orientation and the last a period of specialization.

Entry into a higher education institution is achieved on the basis of a passing-out examination after the completion of a full six-year course and a subsequent special examination in certain subjects. The former is internally administered by the school but controlled by a special jury to ensure uniformity of standards throughout the country (something extremely difficult, given the multilingual, and subsequently multicultural, character of Belgian society). Successful candidates are then awarded their certificat d'humanites which confers on them the right to present themselves for an examen de maturite in three subjects related to the field of study they wishes to pursue at the university.

In the Netherlands, the development of a "pluralistic" system of educational instruction has been striking, because it comprises a multitude of sub-types within each type of educational establishment. More precisely, the pre-university general education is subdivided into four types of secondary school: a) the first, known as VWO, covers the ages 12-18 and consists of three kinds of schools: gymnasium, athenaeum, and the integrated VWO; b) the "senior general secondary education," known as HAVO, covers the ages 12- 17, and is primarily designed to prepare pupils for higher vocational education; c) the "junior general secondary education" (MAVO) covers the 12-16 age range, and prepares pupils for the b type; d) finally there is also the "elementary general secondary education" (LAVO), which used to cover the 12-14 age range, but over the years have been absorbed into larger combined schools.

Higher education is itself fragmented into numerous establishments, representing a wide diversity of school types. Thus, while university entry is possible only after completion of the six-year "pre-university" course (VWO), admission to other higher education institutes is based on the successful completion of a cycle of studies in a relevant technical-vocational secondary school, although the students are required (the same applies to Belgium) to "have an adequate knowledge of the basic subjects: mathematics, physics, chemistry and biology" (ibid., p.214).

3. The French System

A tension between the individualist and collectivist strands in French educational ideologies can be traced in the concepts of the famous revolutionary slogan "Liberty, Equality and Fraternity." With the passing of years one could claim that the first two concepts receded in favor of the third.

The 1975 injection (the so-called "Habby Reform") that education "should prepare children for working life" can be seen in the context of the transformation of France from an agricultural society into one of the most advanced industrial countries in the world.(Note 2) Since the end of the Second World War, but particularly after of the Gaullist government in 1958, a manpower planning approach has influenced the establishment of educational priorities. The growing importance of individual rights justifications for education in the post-war period has been tempered by the priority given to an economic society-centred aim.

Lower secondary schooling (in the form of the college d' enseignement secondaire, or CES) became available to all children in 1959, although there was also a restriction in that the admission to post-primary education was decided on the strength of the primary school records, and on a type of examination by a commission which included parental representation. Strict examinations led (and still do lead) to the upper three-year cycle (class de seconde). Those who succeed in entering this cycle are bound after three years to sit one of the various baccalaureat examinations which lead to university training. Those who fail follow a "short" course of further training. However, one of the major changes in the *lycees* since the 1960s has been that branches of the baccalaureat have been introduced that have a technological orientation alongside the traditional academic course.

Increased participation in higher education has therefore been achieved through the creation of "lower standard" Baccalaureates which have resulted in more of the less academically able students entering higher education courses, which they find too difficult to complete. When they fail, the system provides them with an opportunity to either retake the year or change their course of study; this raises the cost of their education to all stakeholders: the students, their families, employers and the nation as a whole. Formally, selection for higher education has been ruled out but there is already a significant amount of "unacknowledged selection" which takes place at the time of admission and by later examinations to ease the strain on resources, especially within the first two years.

In higher education, reforms after the 1968 social unrest--and the subsequent feverish debates it generated --made bold steps towards a system that could secure more autonomy to universities, reduction of the privileges of certain academic faculties, broadening of studies, and student participation in university government.

Access to the university has been widened to include students from more "disadvantaged" backgrounds. Central allocation of resources has reduced the geographical inequalities found in some other educational systems. Even if someone argued that the position of certain establishments (i.e. the Grandes Ecoles) "has been little threatened by educational reforms" since "the elite of French society was educated outside the mainstream university system" (McLean, in Holmes, 1985, p.91), we should not forget that the prestige of such establishments is being tested everyday in the highly competitive system of a "global market" of higher education services.

4. The German System

Until the beginning of the 1960s, the extension of state activity in the direction of an "active intervention" faced strong opposition in the Federal Republic of Germany so that planning was not then an issue. At the same time, we should not forget that the responsibility for the school system--according to the Basic Law of 1949--does not lie with the federal government. The individual federal states (Laender) are independent in educational and cultural matters.

What strikes the observer of the German system--at least the western part in the pre-unification period--is the emphasis placed on the "manpower" approach to the design of education, i.e., a deep concern for the alignment of the school structure with labour market demands.

However, increasing foreign competition in industrial products, an unprecedented flow of immigration from low-income countries during the last two decades and the shock of unification in 1990--and the enormous costs it has entailed--have caused a crisis not only in the social welfare system of the country, but they have brought into question the effectiveness of the vocational system itself, given the high rate of unemployment (at about 11% in the late 1995). In addition, the employers, hitherto very supportive of the apprenticeship system, started to complain that in the wake of technological change, they would need more less-skilled but flexible workers, able to switch easily from one task to another.(see "The Economist," 6/4/1996, p.23; also 4/5/1996, pp. 11-12 and 21-23).

Admission to universities is made on the basis of the secondary school leaving certificate (*Abitur*), as well as the university-entry examinations (Hochschulreife). This system of higher education is highly differentiated. Classical universities compete with colleges of advanced technology and teacher-training colleges, as well as with private universities and technical colleges. There are no tuition fees at German universities or tertiary colleges. Here it must be noted that an increase in the demand for higher education during the 1960s and 1970s caused the creation of 13 new universities, which not only satisfied the public pressure for higher education, but in addition introduced a number of innovative schemes. An example was the creation of an experimental comprehensive school attached to the university of Bielefeld (founded in 1967) in order to serve as a preparatory stage for the first year of university study. (Mallinson, 1980, p.235).

A system relatively similar to Germany's has been developed in Italy, and in Switzerland as well. Especially as far as technico-vocational education is concerned, arrangements have been made to promote a sound basis for large-scale industrial development.(Note 3)

As a result of centrally initiated efforts--albeit with the full co-operation of the regional administrations--technical and vocational education have enjoyed comparatively favoured treatment. Of particular interest is the fact that, as Mallinson (1980) revealed, in the early 1980s from the graduates of the schuola media (the four-year compulsory "intermediate" school, following the five-year "elementary" school), 34% enter the five-year instituto technico, which awards his/her holder with a "mature diploma," enabling him/her to "either go directly to some form of tertiary education at the university level, or to enter into higher grades of management" (p.248). However, we should note that this kind of school--and indeed any other type of higher secondary (non-compulsory) school--charges fees, but the inequalities are not so profound since this handicap can be minimised from the attachment of "equal status" and relatively equal opportunities for access to higher education to all the pathways.

In the Italian model the concept of "higher education" is virtually identical with that of "university." During the last 30 years, the intake of the universities has been increased dramatically. This is due not only to the increased internal demand for higher education (after all, the proportion of the higher secondary school graduates who register in universities remained quite low, at 30%) but also to an impressive inflow of foreign students, who were encouraged by the rather loose criteria for admissions, especially as far as the E.C. citizens are concerned. Thus, the number of students entering university is still increasing at an estimated rate of about 27% between 1986 and 1990. (Brock & Tulasiewicz, 1994, p.172; also UNESCO, Statistical Yearbooks of respective years)

5. The Iberian System

In the educational systems of Spain and Portugal, there are clear boundaries between the different paths of secondary schooling, as well as between the primary and the secondary level. Basic education in both of these countries is free of charge and compulsory, and lasts for nine years. Secondary education is divided into academic and vocational branches without any interconnection between them. In addition, there are many private schools run mainly by the Church, which-- given the high religious solidarity characterising the Catholics--plays a very important role in the educational policy-making.

Although Spain has a more decentralised administrative structure (it is divided into 17 autonomous communities), one could claim that both of these two countries "designated" the state to ensure the basic unity of education and guarantee equal conditions for all in the exercise of their rights.

In Portugal, university entry is achieved through the leaving-certificate of the general secondary school, whereas in Spain there is a transitional year linking school and university, at the end of which students are evaluated and then a university entrance examination is taken, more commonly known as Selectividad. (Brock & Tulasiewicz, 1994, pp.244-246 and 264-265)

The percentage of those gaining access to university education in Portugal is quite small, and far below the European Community average, although there are clear governmental short and long-term objectives for balancing the disparities in the distribution of students in regional institutions, promotion of short-cycle polytechnic education and compensatory measures for the underprivileged students. (ibid., p.246) In Spain the participation rates are far larger (some 78% in 1987/88), although there is a system of university fees, depending on the course.

6. England and Wales

The education system of England and Wales has witnessed many radical changes in the last fifty years, especially in the administration of the schools, the curriculum content in the state-maintained secondary schools and the post-compulsory schooling options. (There are certain features in the administration and structure of the Scottish and Northern-Irish systems that do not justify consideration of the system of the United Kingdom as a whole.)

The entire structure of secondary and post-secondary schooling has been repeatedly revised, and numerous experiments took place during the 1950s and 1960s, either in at the national or local level, and included the state-run (public funded under the administration and supervision of the local authorities) as well as the privately-run establishments. In the whole controversy about the structure of the system, political and ideological claims as contradictory as those for "equality of opportunity," on the one hand and "high standards" on the other, have been presented in the agenda. Initiatives--not always derived from purely educational considerations--have been taken in different places of the country, due to the decentralised nature of the educational decision-making as well as the lack of a nation-wide consensus on what constitutes an appropriate secondary schooling.

From the mid-1960s onwards, the dominant type of secondary school became the so-called comprehensive school, albeit with great variations and, most of all numerous implementation problems. This kind of "integrated" school, which combined elements of traditional academically oriented curriculum as well as vocational instruction, was persistently under attack, especially during the 1970s, when the global economic crisis caused deep concern about the "effectiveness" of the system in a world of undeniable financial stringency.

The introduction of a "National Curriculum" in 1988, created "core" and "foundation" subjects, in relation to which each pupil in state schools would be expected to have a certain amount of "knowledge", "skills" and "understanding" at the end of pre-specified age-related levels (key stages). (See DES, 1988, section 2) At the same time, standardised assessment practices--under the auspices and encouragement of the central government-- were proposed and tested in various experimental programs all over the country.

As far as the parallel to the general secondary and post-secondary schooling system is concerned, the focus of the state policies has been on the enhancement of further education, which has increasingly been seen as embracing the 14-18 age group. Introduced from the 1970s onwards were the vocational certificates like the GNVQ, the HNC, the HND, and many more, awarded not only by the state and the local authorities but by independent professional bodies as well. In addition, an influx of new schemes such as the Youth Opportunities Program (YOP), the Youth Training Scheme (YTS) and the Technical and Vocational Initiative (TVEI) led "to schools being funded from sources with very clear strings of an instrumental and vocational kind, and thus to a shift of emphasis within the education system back to that of the "industrial trainers"" (Kelly, 1990, p.39).

The distinction between the old GCE (General Certificate of Education) O level and the CSE (Certificate of Secondary Education) survived until the 1980s when these two types were integrated into the GCSE (General Certificate of Secondary Education) O level, after widespread criticism for inefficiency. However, the GCE A level examinations remained the number one factor affecting university admissions, with new subjects added during the last few years, which quite often constitute an inter-disciplinary approach in the various areas of knowledge. That, in combination with the establishment of new universities (among which there is a number of the former Polytechnics, known in the past as Colleges of Advanced Technology) led to an expansion of higher education, something that is in a reverse direction lately, since the government funding--that covers over 90% of the universities recurrent and capital expenditures--has been dramatically reduced.

General Trends of Selection in Europe

Although there is no single pattern of selection procedures throughout Western Europe--not to mention the former "socialist" countries of Eastern Europe--it is possible to trace in the systems examined above certain characteristics that reveal common elements and mechanisms, which, far from constituting a starting point towards a "policy of harmonisation" (Mallinson, 1980, chap. 10), at least offer a comparative view of the context in which the Greek system develops and of the influences that are being exercised upon its structure.

I shall refer extensively to the Greek educational system; it will be noted, for example, that these systems offer more flexible arrangements in the school life of children, better developed branches of vocational and technical training and wider opportunities for adult and continuous education. This is not to argue that these systems are heading towards a more "democratic" and "equal opportunities" future. Despite politically coloured declarations about "educational provision for everyone" under an environment that "favours the individual's aspiration" and "respects his/her socio- economic background", we must not forget that in a world of global competition and market domination, concepts such as "inequality" and "social justice" give way to the notions of "individual success", "value for money" and "monitoring of standards".

As many international studies have shown, in some countries an "equalisation" among socio-economic strata has emerged, while in others virtual stability is the case. For example, in a comparative study made by Shavit (1989), in countries like Germany, Switzerland and Sweden, the expansion of secondary education has been accompanied by a growing differentiation into academic and vocational tracks or programs. The expansion of vocational, non-college education enabled these systems to incorporate a growing proportion of the lower strata who would complete secondary education but would not be considered for further academic education. As a consequence, they have witnessed an opening up of secondary education without disturbing the basically exclusive character of higher education (see also Shavit & Blossfeld, 1993, pp.20-22).

The general pattern observed in some of the comparative studies that examined patterns of selection in Europe is that, as successive generations go through the education filter, the proportion of those gaining a place in a level which would have been inconceivable two or three generations before has been considerably increased, especially in the lower socioeconomic strata. The effects of social origins are generally stronger "at the beginning of the educational career and then decline for subsequent educational transitions" (Shavit & Blossfeld, op.cit., p.18). These findings relate more to the so- called "life-course" hypothesis, which states that "if primary and lower secondary education become universal and lead to a decrease in the effect of social origin at these earlier levels, then the effects of social origin on higher grade progression will stay small across cohorts because older pupils are less dependent on the preferences and the economic conditions of their families than younger ones" (ibid., p.9).

Far from suggesting that there has been a drastical reduction in the association between social origins and any of the educational transitions, this presents a trend in highly developed (post)industrial western countries showing that inequalities in the transition stages throughout the various education levels have been progressively more complicated than before. Whereas in the past it was relatively easy to define what the class boundaries were, or which particular types of educational instruction were the more prestigious ones, today influences other than socio-economic status (traditionally measured as the parents', and specifically, the father's occupation) contribute to the opportunities of "success" (a term rather subjective in itself). Such other variables include the "attainment of private tuition classes", the "multiplication of scientific disciplines in higher education", the "changing status that different professions have in a rapidly advanced modern society", the "emergence of new youth cultural stereotypes" etc.

In addition, in countries, such as the Netherlands, or Sweden, the existence of so many vocational paths parallel to the general education network, has offered a widening of opportunities for those hitherto deprived of access to modes of further qualification to enrol in a course with promising future prospects. Some research evidence maintains that there exists a considerable decline in the effect of social origin on educational attainment, in these countries.(Shavit & Blossfeld, op.cit., chap. 5)

But clearly this is not the case in the overall picture of the educational opportunities in Europe. While there has been a slight narrowing in rates of participation, the proportion of students from higher socio-economic backgrounds has not changed radically, especially in higher education, which is the most important (given the opening up of access in the lower levels) level at which to measure the persistency of inequalities. More importantly, when increases are recorded in the participation rates of students from the "disadvantaged social groups", they tend to occur mainly "in the less prestigious programs of the higher education sector".

There are, of course, variations that stem not only from the university admissions policies, but also from the structure and organisation of the secondary school. In systems with a tradition of "open access" to higher education, on the basis of minimum educational qualifications (e.g., France, Germany) the selection procedure starts very early in children's school life with a very elaborate selection mechanism that sorts out progressively the "best performers" in academic education. If one takes into account the financial squeeze of the recent years in these systems, then one realises that in practice the "open system" policy is being progressively replaced by the introduction of a numerous clausus provision, which has the effect of making students compete for entry.

In systems where the selection is made on the basis of scholastic achievement (e.g., Sweden, Spain and partly Britain), the selection is made relatively late, and there is also a tendency for standardised and externally administered procedures of assessment. (Christie and Forest, 1981) This approach to selection (reliance solely on performance on an external examination) as a basis for school certification is increasingly being questioned, not only on the ground of its unsuitability for assessing individual needs, interests as well as varied curriculum areas (Ball, 1990; Kelly, 1990), but also its "failure" to change the prevailing pattern, namely, that "those whose fathers were highly educated and had high prestige jobs more often obtained tertiary qualifications". (Kerckhoff and Trott, in Shavit and Blossfeld, 1993, p.151; also Halsey et al., 1980)

On the other hand, a entirely school-based assessment, despite its suitability of involving a wide sampling of student achievements, is not very popular in most of the examined countries, because of comparability problems arising specifically when the results are used for qualification or selection purposes in highly competitive labour markets.

With the above evidence and considerations in mind, it is difficult to distinguish a single selection system, and characterise it as more or less "fair". In response to the needs of the changing workplace, as well as to accommodate the needs of students, curricula contents and school structures are already changing in secondary and tertiary institutions. It will be a daunting task to devise selection procedures that do not have serious negative impact, not only on teaching and learning in schools, but in the social differentiation.

The introduction of modern methods of teaching has not yet brought any significant change to the opportunities for access to higher education. The relatively "open" school environment has remained a "privilege" of the younger age- cohorts at a time when the selection processes, considered as vital in European educational systems-- in contrast to the American system consisting of the "shopping-mall" schools-- preserved their exclusiveness only at or nearly at the end of the secondary school.

Despite the progress so far made in the school systems around the world with the introduction of new practices and modern pedagogical methods based on group-learning principles, really noticeable changes have a long way to go to be adopted, especially in the upper levels of secondary schooling. In other words, although the curricular structure and the pedagogical framework at the lower levels can be quite "elaborated" (to use a "bernsteinian" term), at the upper levels only individual effort and performance are rewarded. That becomes clearer and more decisive at the transition period between secondary and tertiary education. Thus, one could argue that at levels of the educational "ladder" where the degree of specialisation and the need for individual selection are insignificant, collective learning flourishes; on the contrary, at the upper levels, and especially as the time to enter the labour market approaches, the student's success depends entirely on his or her school (and examination) performance, and collective learning disappears.

A number of "technocratic" solutions supported in recent years by policy-makers, either in the direction of "maintaining high standards" (e.g., introduction of numerus clausus policies), or of "comparability" of the various systems (e.g., use of "standardised" testing methods), present selective admissions as a "need" and, most of all, as an objectively assessed procedure, according to universally accepted principles. The role that social factors (such as class interests) play in the definition of a "worthwhile corpus of knowledge" is being continuously undermined.

Selection in Greece

The 1964 reform.

The reforms in examinations and curricula came as a response to the general climate of political freedom and democratic changes, and the feeling that the formal educational system was obsolete and maladjusted in relation to the context of a rapidly changed, and technologically advanced "western world". Not surprisingly, the general system of schooling, especially in its basic levels (primary and secondary) was the first target of the reform policies of the various governments.

Although the most fundamental changes in the system were introduced after the re-establishment of democracy in 1974, the "seeds" for these changes had existed in the mid-sixties, when the liberal party of George Papandreou was in office. Among the changes brought by this government--which attempted in an unfavourable social climate to attack the "classicism", "conservatism" and "intellectualism" of the preceding right-wing governments --the most important were the following:

- free education at all levels of public education
- nine-year compulsory attendance instead of the six-year system existed
- restructuring of secondary schools into three-year gymnasium

(lower) and three-year *lyceum* (upper); the latter would include general and technical-vocational types.

- "demotiki", the vernacular, would be the language of instruction in primary schools and taught along with "katharevousa" (a simplified form of ancient Greek) in secondary schools.
- at the end of secondary education the pupils would sit special examinations to get the "academic certificate" that would allow entry to the universities.

As far as the last change is concerned, one can argue that it was actually the first official step towards the introduction of a National Examinations System. In other words, whereas in the past each higher education institute had been conducting its own entry examinations, now the Ministry of Education was responsible for these examinations at the national level. The examinations for the "academic certificate" (something analogous to the French baccalaureat and the German *abitur*) were to be conducted on specific dates in various cities of the country, based on the subject matter taught in the *lycea* and marked by secondary school teachers, not university professors.

In order for a candidate to get the "academic certificate," he or she had two alternative "types of schools" to choose from. The first one included the so-called "liberal" disciplines (Law, Literature, Theology, Economics, Political Science and Teacher-training Colleges), whereas the second one included "applied" disciplines (Natural Science, Mathematics, Medical Science, Engineering, Architecture, etc.). However, even the latter were affected by the traditional orientation of the whole system, since among the examined subjects were "Ancient Greek", "Modern Greek" and "History", although in the calculation of the final grade they were multiplied by different coefficients than those of the type A subjects.

It was not the first time that a nation-wide educational reform was being attempted. Actually this was the third attempt, in the same century (the previous ones had been made in 1913 and in 1929, both under centre-wing, liberal governments) to insert a new way of thinking into the "conservative" and "classicist" Greek educational structure. The new liaisons between the country and the EEC--from 1961--urged a radical economic restructuring, something that required a equally radical reform of education, which should eventually bear the responsibility to train the necessary work-force. The shift to technical-vocational education was thought capable of creating a multi-leveled network of practically-orientated middle schools that would run parallel to that of the "general" secondary education, enabling Greece to keep pace with the educational systems of the more advanced capitalistic countries. (Bouzakis, 1991:104-105) In addition, measures such as free education at all levels of public education, nine- years' compulsory attendance, the introduction of demotiki, the improvement of teacher-training, etc., generated a sense of "justice" and "equality", especially in the lower social strata where economic and other reasons prevented the provision even of the most basic forms of education.

The introduction of nine years' compulsory schooling raised considerably the participation rates, especially in secondary education throughout the 1960s. For example, overall participation at the secondary level, from 33% of the relative age group in 1960 went up to 56% in 1970. The corresponding figures for girls' participation showed an even more impressive trend, with an increase from 28% to 54%.(OECD, 1980, p.121) In fact, participation of the female population in general secondary education almost equaled that of males-- with slight regional differences. (See Table 1) One could even argue that, on the one hand, girls were more favoured than boys as far as general education was concerned.

Nevertheless, things were reversed when technical- vocational and higher education were included in the estimates. When this was done, the participation of girls in the former, and of lower socioeconomic strata in the latter, was quite low. Especially in the technical schools, girls' level of participation--although the overall enrolments increased dramatically--remained far below that of boys. In 1975- 76 (10 years after the introduction of the reforms),there were only 6,863 girls registered in upper- secondary (middle) technical schools in a total of 55,503 students (12.3%), and 2,607 in lower technical schools in a total of 60,119 (4.3%). In contrast, their participation in vocational schools was significantly higher (middle level: 7,892 in a total of 16,969, or 46.5%; lower level: 1,168 out of 1,308, or 89.2%). (Ministry of Education, 1975, table 2.101) But again, the relatively high proportion of girls in vocational education may be explained by the fact that this type of education, in contrast to the technical type, leads to occupations which are socially accepted as "women's jobs".

Table 1
Participation of Population aged 16-18
in Upper Secondary Education, by Region and Sex
in 1960-61 and 1970-71

	1960-61		1970-71	
	Boys	Girls	Boys	Girls
Region Total	30	23.5	35.8	38
Greater Athens	43.9	42.1	38.2	43.8
Rest of Central Greece & Euboea	24.1	15.5	29.8	30
Peloponnesos	31	25.3	36.9	41.9
Ionian Islands	28.5	21	38.8	31.8
Epirus	29.6	11.1	36.2	32.7
Thessaly	23.8	13.4	40.7	38.2
Macedonia	24.8	19	33.1	36.6
Thrace	13.9	11.6	27	23.5
Aegean Islands	27.3	20.4	39.2	32.1
Crete	29.7	24.6	38.1	42.7

(Source: Ministry of Education, Statistics of Education, 1970-71.)

On the other hand, for higher and university education in the 1960s, the over-representation of social groups very high in the occupational ladder was more than obvious. For example, students whose parents were holders of "professional", or "managerial and administrative" jobs had--in relation to the respective occupations" representation in the whole population--on average from two to four times more chances of securing a place than the sons and daughters of the "blue-collar workers" and the "farmers". (OECD, 1980, p.126). Nevertheless, the gap in the opportunities tended to narrow, not only in socioeconomic, but also in gender and geographic terms, as one moves toward the 1970s. (Polydorides, 1995, chap. 5, 13) This was due to the compensatory measures following the 1964 reform, and the greater State intervention in the reorganisation of all levels of education after the delay caused by the seven- year "break" of the dictatorship, as will be discussed now.

Reforms under the "junta" regime

The 1967 "junta" brought to a halt every reform attempt, and reinforced the conservatism of the "traditionalists". Among the counter reform measures of the period 1967-74 were the reduction of compulsory education from 9 to 6 years, the abolition of translated ancient Greek literature texts, and the replacement of social sciences in the new curriculum. The teaching of "demotiki" was restricted to the first 3 grades of the primary school. Secondary education remained "integrated" in the form of the six-year gymnasium. In general, in those years Greek education was more classics-oriented, bookish and old- fashioned than in the previous decade. The most important changes were to be observed in the "hidden curriculum" (Young, 1971) of the schools and the disciplinary environment in which the teaching was taking place. The regime attempted to turn the attention of the pupils towards the values of the past, especially through the emphasis on ancient Greek and its simplified form of school instruction, the katharevousa (i.e.,the pure language).

The paternalistic attitude of the dictators in "saving" education can be seen in the will of the regime to maintain education free of charge at all levels, and to replace the old textbooks with new ones, aiming at imposing the new Helleno-Christian ideals. (Note 4) A quick review of, say, the textbooks of civil education of that period, would reveal feverish (State-guided) attempts to restore--if it ever existed--the self-confidence of the "nation" through the invocation of the old "virtues" of the glorious "helleno-cristianic past", the condemnation of communism, and unquestioning conformism to the formal guidelines.

Nevertheless, it seems that even in the period 1967- 74 there existed a political will--after insistent recommendations by international organisations like the World Bank--for a restructuring of higher education, on the one hand, and a promotion of technical-vocational education in a higher level, on the other. Thus, although there had been a widespread bias against technical and practical studies in the curriculum-orientation of schools, at the same time it started to be realised that there was a lack of balance in the provision of school knowledge. The output of graduates from secondary schools and universities was growing more rapidly than the capacity of the economy to create new jobs, whereas the output of graduates from technical schools could not meet the shortages in the labour market. It becomes reasonable to assume that the need to increase productivity and improve the overall performance of the economy gradually prevailed over the "liberal" orientation of the Helleno-Christian tradition.

To argue that big improvements have been brought about however, would be naive. Half the population of the country is concentrated in the Greater Athens area, and as a result increased enrolments as a proportion of the respective area-population represented a very small improvement in real figures.

Moreover, if one takes into account the unsystematic methods of collecting data or keeping school records during that period and the high drop-out rates--especially in the rural areas where the contribution of the younger members of the family in agricultural work was considered essential--then it is easier to see that little improvement was achieved in the reduction of regional differences (Eliou, 1976).

In the technical field, while secondary education was marked by the total reverse of the 1964 reform, a new type of educational-development plan emerged under the auspices of foreign guidelines. It was part of a long- term plan of economic development, called "Model for the Long-term Development of Greece, 1972-1987". The Plan projected the mass movement of the labour force from the primary sector (agriculture) of the economy to the secondary (industry) and the tertiary (commerce, services), and in addition an increase in the graduates of technical-vocational education, from 335,000 in 1971 to 1,600,00 in 1987. (Bouzakis, 1986, p.111)

The flow of internal migration toward the big urban centres--taking place in the last three decades--was not the intended outcome of macro-level manpower planning, but rather the result of the complete absence of regional policies throughout the period examined. Moreover, the distribution of the labour force to the secondary and tertiary sectors has been greatly unequal in favour of the second, which reveals the imbalance that characterises not only the productive base of the Greek economy but also the curriculum content of schools and the attention paid to the technical-vocational education. (Note 5)

It is true that the increase in the enrolment and output ratio of technical education during the 1967-174 period was of an unprecedented level for Greece. While the output of the six-year gymnasium increased between 1968 and 1974 at a rate of 37.6%, that of the (lower and middle) technical-vocational schools increased at a rate of 78.8%. The graduates of the latter were about 42% of those of the former (15,898 as compared to 37,844) in 1968, but in 1974 the proportion was 58% (28,657 and 49,183, respectively), although it started to fall again in the following years. (OECD, 1980, p.132) Despite the improvements, the importance given to the "helleno- cristianic tradition" and the mainly classics-oriented curriculum of the Greek schools at that period affected not only the content of technical education and the resources allocated to it by the government, but also its "status" in the eyes of the public. (see Drettakis, 1974; Dimaras, 1975; Noutsos, 1978; Nikta, 1991) As a result, secondary technical-vocational education continued to attract that kind of pupils with no hope of having access to "prestigious" occupations, that is those who were expected to "benefit" by a vocationally-oriented educational provision (blue-collar workers, office clerks, farmers, and the like).

Such was also the case with the participation in higher education by different socioeconomic groups. According to OECD calculations, in the years before the 1976-77 reform, participation in higher education was highly unequal with respect to father's occupation. Although the situation from the 1950s to the 1970s had been changed in favour of the "lower" professions, in the mid-1970s there were still wide differences in access to certain university departments. In 1975-76, for example, Humanities was the only field where all occupational categories were represented almost "equally", whereas Law was "over-represented by professionals and managerial personnel", Social Sciences and Teacher-training were "over-represented by people in agriculture and by blue-collar workers", and the more elite occupations were "concentrated in the more "elite" fields of study, e.g., Medicine". (OECD, 1980:121)

The restructuring of higher education was initiated with the introduction of various "Cycles of Schools" for National Examinations purposes, corresponding to a very "specialised" classification of academic disciplines. (See Figure 1.)

The promotion of technical and vocational education in a higher level was characterised by the establishment - in 1969 - of the "Centres for Technical and Vocational Education", known by the Greek acronym "KATEE". These Centres were created to provide technical education and training for middle-level manpower at the higher technician level, and have been considered as equivalent to the "Community Colleges" in the USA. The major reason of their existence was that they catered for those students "whose initial educational aspirations was a university degree", but "having failed to enter universities, they were obliged to pursue their education elsewhere". (Kalamatianou et al, 1988:272) The first five KATEEs were established in 1974 in Athens, Thessaloniki, Larissa, Patras and Heraclion (Crete), covering 21 faculties with 74 specialised departments.

Figure 1
Examples of Cycles of Schools

Cycles of University Schools	Subjects Examined
A.Literature	Written expression, Ancient Greek, History, Latin
B.Law	Written expression, Ancient Greek, History, Latin
C.Physics-Mathematics	Written expression, Mathematics, Physics, Chemistry
. . .	. . .
H.Economics	Written expression, Mathematics, Geography, History
. . .	. . .
L.Theology	Written expression, Ancient Greek, History, Latin

The 1976-77 Reform: Main Changes

When democracy was re-introduced in 1974, the climate favoured major political, social and educational reform. The recommendations of international bodies, such as The World Bank and OECD, pointed to the great need to support technical education, while they commented on the great "inequalities" in educational opportunities, prevailing in the 1970s, in relation to gender and socioeconomic status. (see OECD, 1980)

In the new formulation of educational policies, what was sought by the governement was no longer the advice of individual Minister-appointees or associates, but rather the establishment of a body of experts who would formally operate as part of the "managerial" group of the State organisational mechanism. (See Parsons, 1960; Blau and Mayer, 1967) The function of this group was perceived as crucial because, not only the complexity of educational innovation in international level and the example of other European countries, but also the need to achieve consensus in such highly controversial reform attempts, required the participation of experts of as many different scientific disciplines as possible. This was contrary to the past when the School of Philosophy of the University of Athens had been the main agent of policy-planning in school matters. The existence of an anti-reformist alliance-- consisting of MPs belonging to the governing right-wing party, individual university professors, appointees of the dictatorship in the State mechanism, and conservative religious pressure-groups--despite causing a delay in the approval of the reform legislation, did not decisively affect the proposed changes.(see Mattheou, 1980, chap.5)

As a consequence of the new policies, in 1975, the Centre of Educational Studies and In-Service Training (KEME) was established. Its main tasks were defined as: "a) the systematic scientific study and research of educational matters, b) advise on any law draft proposal by

80

the Minister, c) the design of textbooks and timetables and d) the in-service training of teachers". (Nikta, 1991, p.63)

Here it should be remembered that the centralised nature of the administrative structure of the Greek system could not - and indeed did not - allow KEME to be involved in essentially political decision-making procedures. Neither did it leave any doubt about the real influence this body of experts had on the educational goals that each political party in power had already set up according to its own ideological principles and political interests. Although it can be asserted that, after 1974 the "political centre" (Ministry of Education) with its affiliated agents--operating under the financial constraints imposed by the government's budgetary policies--was not any longer "impenetrable" from "external" influences (see Archer, 1979), at the same time there is no doubt that all the initiatives for educational reform have been--actually still are--channelled through various patterns of "political manipulation" occurring between the governing- party elite and the different interest groups.(Mattheou, 1980; Eliou, 1986)

One of the main focuses of the debate at that period was the structure, content and orientation of the pre- university level of general education, and especially the undifferentiated general secondary school. The concern with this stage of schooling was decisively influenced by the progressive integration of Greece into a system of international co-operation, particularly after the construction of closer ties with the European Community, of which Greece would become a full member in 1980.

Under the laws 309/1976 (for general education) and 576/1977 (for technical-vocational), secondary education was split into two independent "cycles": the 3-year comprehensive gymnasium (13-15), and the 3-year selective *lyceum* (16-18). The examinations for passage from primary to secondary school were abolished, and strict examinations at the end of gymnasium were introduced in order to allocate pupils to the new diversified senior high school. Those who performed better in these examinations were accepted into the "prestigious" general *lyceum*, whereas those who performed "poorly" were allocated either to the 3-year technical-vocational *lyceum*, or to the 2-year technical-vocational school.

The curriculum of the general *lyceum* was oriented toward higher education. The subjects taught in the first year of this school were common for every pupil. In the second year there was a distinction between common and optional subjects, the latter leading at the end of the third year to one out of two types of certificate (apolytirion), which corresponded to two different groups of academic disciplines. The old classical and practical directions were reshaped into two groups of selectives: a) ancient Greek, history, Latin; and b) mathematics, physics and chemistry.

Ancient Greek was the subject that all students had to attend for the most hours every week. The first of the other two alternatives (the one- to two-year technical-vocational school) was of admittedly lower esteem, and it offered mainly a preparation stage for the labour market, after a kind of specialised training. For those studying in this option, the choice was offered--instead of starting work immediately after graduation--to take special examinations in order to gain a place in the second year of the technical- vocational *lyceum*. The latter was formally considered as having the same status as the general *lyceum*, since it was not only meant to prepare students for the labour market, but in addition it allowed the "ablest" of its graduates to enter higher technical education. For that reason a proportion of the entrees in the KATEEs (32%) was allocated to the technical *lyceum* graduates, according to their school achievement. Thus, we see that whereas entry to higher education (universities, teacher-training institutes and KATEEs) was allowed only after examinations at the national level, the government decided to give a small incentive to those attending technical secondary schools, trying in this way to attract more students who wished to pursue a more vocational type of study, but at the same time were willing to continue their studies at a higher level. (Bouzakis, pp.112-115)

As far as the university-entrance examinations were concerned, the Ministry of Education in 1980 abolished the old system of "Cycles of Schools" and re-introduced the system of two-directions certificate mentioned above. There were two kinds of certificate (the participation in examinations was compulsory in order to graduate): one (Type A) for those wishing to study Humanities, and another for those who chose the "positive" track (Type B). The innovations of this system were the following:

- the subjects examined in the Type B examinations were now more related to the "applied" character of the respective scientific disciplines, in contrast to the previous similar system of 1965 where the dominance of the "liberal" studies was obvious.
- the examined matter was selected solely from the curriculum of the last year"s curriculum.
- the selection of students for universities was made according to the preference of the candidates and their scores. The scores would determine whether a candidate would study in the more prestigious schools within the group that he or she selected.. The score was the weighted mean of the total of grades on their certificate, (second-year achievement + last-year achievement), their grades in composition in modern Greek, and in selective courses at the examinations multiplied by a different component for each school.
- the graduates of technical-vocational *lycea* could sit in the examinations if they had chosen the additional courses of the second (type B) group of electives.

The examinations were called "Panhellenic" (national) because they were taking place simultaneously throughout the country with common subjects selected by a special committee of the Ministry. The examination papers were marked by two secondary school teachers, and in case of a large disparity in their marks, the paper was re-evaluated.

Critical Assessment of the Reforms

The 1976-77 reform did not radically affect the "prestige" of the traditional "academic" subjects. The aim of the *lyceum* was quite similar to that of the upper level of the old 6-year gymnasium, in the sense that it was perceived as a preparatory stage to tertiary education, despite the official declarations that it meant "to provide an education that is richer and broader than that of the gymnasium". (See Law 309/1976, article 29)

In assessing the 1976-77 reform, we must first summarise the major considerations embodied in the laws 309/1976 and 576/1977:

- The raising of the school-leaving age, which was a constitutional mandate (article 16 of the 1975 Greek Constitution), ranked as a very important precondition for the goals of democratisation and modernisation. Compared to other western societies, especially those of the European Economic Community, Greece had the fewest years of compulsory schooling (6 compared to 9 for most other countries).
- Selection through examinations at the end of compulsory schooling (age 15+), and the reorganisation of upper secondary education, would deflate the increasing bulge of aspirants for admission into the universities and other post-secondary institutions. At the same time, they would alleviate the problems of under-employment and psychological frustration.
- Related to the above was a desire to make the education system more efficient and capable of satisfying the economic needs of a "modernising" society.
- A strong wish to maintain control over educational standards, such as the attainment of certain levels of knowledge and the "screening" of the most "talented" for the few places that were--and still are-- of necessity available in the universities.
- The problem of language was bound to be solved, no matter how much delayed that change was. Among educational reformers, the "language question" was not merely an issue over what form of Greek should be taught in school. It represented basic differences in Greek social and educational philosophy. The introduction of the modern Greek language would help open up new cultural and intellectual horizons, those grounded in the contemporary socioeconomic needs of Greece; it would arouse pupils' interest in learning; and, ultimately, it would develop more versatile, responsible and democratic citizens.

Despite the declarations, the general *lyceum* kept the role of training the pupils only for the universities and providing general culture without any consideration for the labour market. The "shadow" of the entrance examinations to *lycea* affected the study of pupils and the curriculum balance in favour of modern Greek, mathematics, history and physics, that were conducive to their success in passing the examinations.

The arguments in favour of the economic benefit of education were applied only in the case of technical education. In this direction there have been relatively rapid changes. First of all, law 576/1977 abolished all the lower-secondary technical schools as uneconomic and unpopular and set the priorities for an extensive program of building construction throughout the country to meet the needs of a sound technical education provision. Thus, in parallel to the general *lyceum*, there was the technical *lyceum*--which, officially, granted its graduates the same status - and the technical school, representing earlier types of technical provision. The latter's existence revealed the financial stringency within which the reform attempts had to be implemented.

The main obstacle to vocational education was the reluctance of parents to accept non-traditional orientations for their children and the reluctance of pupils to abandon their dreams for a job in the public sector. Such attitudes were deeply rooted in their minds since, from as early as the 19th century, the school certificate (and later on the academic degree) had been inextricably linked to a kind of occupational security and social success. (Tsoukalas, 1977)

In general, it can be said that the reform efforts in the secondary stage, despite the big improvements they brought, lacked two things: a) proper timing, in the sense that the State authorities tried after considerable delay to implement a number of changes, many of which could and should have been initiated decades ago (Bouzakis, 1986, pp.121-23), and b) the existence of adequate infrastructure and resources that could effectively support a "shift" to technical education. The conditions for the success of the new system were far below the very promising official rhetoric, especially as far as the training and continuous support of teachers responsible for teaching a new vocationally-oriented curriculum was concerned. The overwhelming majority of the three-year gymnasium still preferred the more prestigious *lyceum* path in even higher proportions than those witnessed heretofore. Thus, in the school-year 1976-77, 93.5% of the gymnasium graduates participated in the qualifying examinations to *lyceum* (general or technical) instead of applying for a place in the two-year technical-vocational school; in 1977-78, the figure was 97.2%(ibid., p.122)

In higher education, the most important aspect of the new system was the "homogeneity" that it brought, at least as far as the characteristics of the student population in each institute are concerned. More specifically, the abolition of the previous "cycles of schools", and the grouping of different disciplines resulted in:

- a very high proportion of candidates entering academic departments completely irrelevant to those that had been their initial choice; therefore many of them either were unsatisfied with their studies or decided to sit again the next year for the National Examinations;
- there seemed to be a kind of "social mixing" going on in the various academic departments, in the sense that now students coming from "lower" socioeconomic strata (e.g. the offspring of manual workers and peasants) gained-- often accidentally--a place to the prestigious disciplines of Technology, Law and Medicine, whereas in the past they were forced to choose one particular group ("cycle") of academic departments. (Papadimitriou, 1991, pp.120-123) In other words, a kind of "equalising" mechanism appeared to affect the distribution of the higher education places, especially in the universities, by allowing candidates with socially "inferior" backgrounds to attend courses previously dominated by the more "privileged" groups (civil servants, self-employed professionals etc.), although the reverse did not happen.

It is important here to stress the selection and credentialling role which examinations were called to perform in Greek education, in a period when the high social demand for general education and for university degrees was running against every attempt to control educational standards and output. Evidence of this demand was the increasing proportion of the entrees in the late seventies and early eighties--although it remained relatively low. In 1974, the number of entrees in higher education (Universities and KATEEs) was 16,025 in a total of 68,063 applicants (23.5%); in 1978, 21,375 out of 84,417 (25.3%); in 1981, 26,754 out of 75,206 (35.5%); and in 1982, 33,235 out of 78,708 (42.3%). (Katsikas, 1994:136) The need to screen the "intellectually most capable" for the few places that were of necessity available in the universities and other higher institutions (the prevailing principle was that of "meritocracy") was indisputable, but at the same time the highly selective procedure through which this was being achieved raised questions of "equity" and "social justice". Especially, educational opportunity within higher education was found to "favour" students with fathers at the highest levels of the occupational pyramid, and that was a conclusion derived not only by Greek researchers (Eliou, 1976; Drettakis, 1979; Milonas, 1982; Fragoudaki, 1985; Polydorides, 1984 and 1986), but also by international organisations like the OECD (1980).

Changes in the 1980s

Social and Educational Context

After the general elections in 1981, PASOK, the socialist party of the opposition--formed only 7 years earlier--came into the office.

One must not forget that the new government came to power in a period of radical sociopolitical and economic transformations. Apart from the first signs of the collapse of the "Eastern Block" in the mid- and late-eighties, Western Europe was experiencing an unprecedented movement for economic integration, through the EEC and its expansion southwards. Greece, as a new member of the EEC was being tantalised by problems of imbalances in the structure of the various sectors of the economy. Despite the fact that Greek per capita GDP jumped from one-quarter of the OECD average in the mid-fifties to almost one-half in 1979, it continued to have a narrow industrial base and a large inefficient agricultural sector, which accounted for 18% of GDP and 30% of employment in 1980. (OECD, 1993:14) At the same time, the country faced--due to the promising industrial growth of the sixties and early seventies--a dynamic internal migration. The urban population, from 43% of the total in 1961, went up to 58% in 1981. A dramatic shift in the distribution of the labour force showed a progressive decline in the primary (agricultural) sector,a decline, however, which favoured the non-productive tertiary (service) sector instead of the secondary one (industry and construction). The distribution of the labour force in 1961 was 53.8% in the primary sector, 19% in the secondary and 27.2% in the tertiary one; in 1981, the figures were 30.7%, 29% and 40.3%. (NSSG, 1961, 1981)

The general context of the policy of PASOK government was aimed at "national independence, the sovereignty of people, social liberation and the socialist transformation of society". (PASOK, 1981, p.13) Among the measures taken by the PASOK government was the reinforcement of compensatory education. A new institution of post-*lyceum* preparatory centres was established aimed at offering free training for the examinations on the selected subjects for all pupils. The aim of these public institutions was to provide tuition to the poor pupils who could not afford private tuition, and to reduce the inequalities of the private cost of exam preparation, especially in rural areas.

From as early as 1982, the entrance examinations to *lyceum* were abolished, and a new type of *lyceum* was introduced in 1984 at the post-compulsory level. This type was called "multilateral" *lyceum*, and it combined characteristics of the general and technical-vocational types. The direction of this school was the integration of general and vocational education, and the elimination of the "prejudice against manual work," the "offering of scientific and technological knowledge, and the methodology of acquiring this knowledge" and the "offering of equal opportunities to all young people, and helping pupils to become democratic citizens..." (Ministry of Education, 1987, pp.19-22).

The pupils in these schools had--and still have--a compulsory core curriculum at the first grade, with very little room for electives. At the second grade, they can choose one out of six "cycles" that are connected to some of the seventeen specialisations at the third grade. That means that one "cycle" in the second grade leads to various specialisations in the third grade. For example, Cycle 1: Man and Society leads to the following specialisations: either academic option 3 to the university faculties of Humanities and Law, or other vocational specialisations such as office tasks, librarians, computing, social services and applied arts". (Nikta, 1991, p.241) Thus, the options are built up and developed along with the grades. The fact that only at the third grade there are vocational specialisations similar to those of technical-vocational *lycea* justified the additional fourth year of practice in the specialities of some lines, such as agriculture, secretarial, car engineering and the like. (Ministry of Education, 1984, pp.839-46) The statistics, however, show that in 1989 only 400 out of 6,130 graduates (of multilateral schools) were attending the fourth year. (Ministry of Education, 1991)

On the whole, the curriculum of this *lyceum* rather more resembles the curriculum of the general *lyceum*, with few additional subjects, than that of the technical- vocational curriculum. It seems--although there are no specific surveys of the balance between theory and practice in the vocational courses of these schools--that the new *lyceum* is rather an improved version of the general (academic) *lyceum*, with pre-vocational subjects at the first two grades, and theoretical vocational training at the final grade.

The technical schools, on the other hand remained "low- prestige" institutions throughout the eighties. According to various reports by the School Advisers (former Inspectors), technical education evidenced an "overlap between the specialisations" offered at the two-year schools and three-year *lycea*, the "old fashioned curriculum content" that lacked connection with production, and the complete absence of visiting or training of the students in industrial environment, so that the existing programs "do not prepare them for the labour market". (Nikta, 1991, p.108) These schools are mainly attended by boys, since the social division for women and men is still very strong, despite numerous campaigns initiated by the socialist government, during the eighties, to bring about gender equality. Kokos (1987) argued that the technical *lyceum* is the "refuge" of pupils from non-privileged social strata, who are forced to seek employment after 18.

The Greek experience showed that not only is the government to be blamed for this situation becuase of its reluctance to design better strategies and pay the real cost of vocational education, but also blame can be attributed to the high expectations that Greek families have towards general secondary and higher education. Job insecurity, exploitation by the employers, poor payment, bad conditions of work and the like, are societal factors that greatly affect their decisions, and credit general education with the highest prestige.

Examinations

The examination system itself has been changed in many ways. In 1983, a new "four-track" system was introduced; and, in 1988, the higher education entrance examinations were separated from the graduate certificate; in other words, pupils were no longer obliged to sit the examinations in order to graduate.

Those wishing to sit the examinations had--and still have--to attend one of four "tracks" (groups of specialisation):

1. the first one leads to university departments of Science and Technology and higher technological institutes, and the examined subjects are composition in modern Greek, mathematics, physics and chemistry.
2. the second one leads to medical and biological departments, and comprises the subjects of composition in modern Greek, physics, chemistry and biology.
3. the third one offers opportunities for entrance to departments like Philosophy, Law, Modern-ancient Literature, Education , and the subjects examined are composition in modern Greek, ancient Greek literature and language, latin and history.
4. the fourth branch leads to departments like Political Science, Economics, Administration, Sociology, and includes the subjects of composition in modern Greek, mathematics, history and sociology.

The new system did not wipe out the examinations to tertiary level as the socialist party had promised. It, nevertheless, offered a more rational distribution of higher education and a greater variety of channels as well as the chance of limitless attempts for the candidates. It also eliminated the stress of these crucial examinations from the secondary grade of *lyceum*, since from 1988 in the calculation of the final score of the national examinations no more have the results of the first and second grade of *lyceum* to be taken into account, whereas in the period 1983-87 these results accounted for 25% of the total score. Thus, in the university-entrance examinations ("Genikes Exetasis", as they were named), each subject is examined on one and only day of the year, which is predetermined by the Ministry of Education. This change separated completely-- typically, because in essence the links have remained unbreakable--the performance of the students in high school and the national examinations, which caused criticism because it does not provide any incentive for greater school achievement, while it exposes the whole process of the assessment to various accidental factors (psychological stress, memorisation, luck, technical problems). On the other hand, the justification for that decision, as the government argued, was based on the effect that the examination process had on the curriculum and its (internal) assessment within the schools; it was argued more or less that a "distorted" kind of competition had been going on in the school classes between the students, something that also raised questions about "commercialisation" of the assessment system.

In the meantime, candidates competing for a place in the university schools increased three times between 1974 and 1986, while the number of those successful in entering only doubled.

The introduction in 1983 of the three-year Technological Education Institutes (TEIs), which replaced the KATEE, has been seen not only as an attempt to improve standards in the provision of higher technical and vocational knowledge, but also as a way of diminishing the trend toward greater competition in the university examinations. The reduction of chances and the stiff competition for university places forced the less successful applicants to turn towards the TEIs. While university has become more inaccessible for the majority of the secondary school graduates, the number of students entering TEIs has continued to increase.

The most important differences between TEIs and universities derive from their officially stated educational objectives. The TEIs aim to "provide education in the classroom and in the "real world" (laboratories, business, experimental fields, organisations and other public or private establishments linked with the TEIs) for technologists". (Kalamatianos et al, 1988, p.273) To achieve these objectives, TEIs run programs which lead to a first degree ("ptychio") after at least six semesters of classroom instruction, plus one or two additional semesters of practical training. Universities, on the other hand, offer programs which lead to a first--but not necessarily final-- degree (also called "ptychio") after eight semesters for all departments, except for engineering and dentistry which require 10 semesters, and for medicine which requires 12 semesters.

Despite the popularity that TEIs have gained during the last decade, there are still problems of "equitable" distribution of higher education places, on the one hand because the percentage of candidates accepted has remained low (about 18% of the total for universities, and 17% for TEIs), and on the other hand because many people still consider university places as "highly prestigious" in relation to the TEIs, and the allocation of their places "is very inequitable and favours high income groups" (Psacharopoulos & Papas, 1987 and 1993).

Selection and Educational Changes

The education reform attempts, as we have seen, have not been quite as bold and radical as one would expect in a country which, officially at least, belongs to the "West", and is considered as an "upper middle-income economy" by all the major international organisations. The move from goals and legislation to school outcomes was blocked by political inertia and the fixed processes of Greek school in a period of decreasing economic growth and austere budgets. The constraints of the inherited institutions and the goal conflict between equality of opportunity and excellence should not be ignored. Despite the relative success of the reforms in expanding secondary education, the Greek schools are still plagued by drop-out rates, the lack of fit between curriculum and job requirements and inequalities in the distribution of educational opportunities.

Many of the innovations brought (mainly) by the governments of the 1970s and the 1980s did not create the necessary consensus among professional groups, political parties and education planners. The main innovation of the PASOK government, the "multilateral school", has not yet been expanded as expected, mainly because of the enormous financial and technical resources required, in a period of very stringent fiscal policies, implemented by the State in order to meet the criteria for the monetary unification of the European Union.

In addition, the monopoly on decision making by the top hierarchy of the Ministry of Education was--as it has been proved by this experience--the primary goal of the policy makers, and this was the framework in which the reform was initiated. Examples such as the central control of the content of the textbooks--and even of the teacher"s manuals--or the changes of the educational hierarchies and priorities according to the changes of Minister, are very revealing of the intentions of past governments.

The historically established access of wide strata of the Greek population to university education has been hindered by the numerous clausus policy. It is important to stress the shift of public pressure for "equal opportunities" in access in educational provision, from the lower to the higher levels of the system. These changes ran parallel to general reforms in the organisation of schools, and responded to: i) a climate of political freedom and democratic changes that favoured a more "egalitarian" school structure, and ii) demands for alignment of the school functions with the needs of the developing Greek economy,i.e., the promotion and support of technical-vocational education, which has been neglected because of the highly selective examination system, and a dominant "academic" curriculum.

The old "classical" and "practical" orientations were reshaped in an effort to keep up with the radical changes in the socioeconomic conditions and technological advancements experienced throughout the so- called "developed world", part of which Greece has always struggled to be. The truth is that the Greek educational system has been dominated from the very beginning of the existence of the modern Greek State by a highly centralised structure, which enabled the governments and the various politically influential groups to impose organisational and financial restrictions on the "democratisation" of decision- making. At the same time, the orientation of the curriculum towards the "classical tradition", obviated for a very long time any possibility for introduction of broader and more balanced content based on modern pedagogical methods.

Revealing of the conservatism characterising the curriculum policies during the last half century has been the prominent and persistent involvement of the School of Philosophy of the University of Athens in the planning and implementation of every major curricular change. Its members have repeatedly opposed every attempt at reform that would "deviate" from the "humanistic" type of education, prevailing in the Greek schools for decades. (Dimaras, 1975; Mattheou, 1980) The dominant ideology has always been that of "meritocracy"; the "few", the most "capable", the most "intelligent" could and should have access to the highest educational levels possible, and enjoy the privileges and social status that the most "prestigious" academic disciplines can secure.

The lack of adequate provision for technical and vocational education was only an indication of the wider weaknesses of the system, and the inequalities (re)produced by its existence. The very restricted access to higher levels of secondary, and later on university education must not be attributed only to the hostile attitude of the "traditionalist" policy-makers toward the reforms, nor to the financial stringency and lack of resources; it must be searched for in the deep-rooted cultural values and principles of Greek society, which, subsequently, may reveal the relation between the dependent character of the country's economic structure and the sociopolitical context of its recent history. The educational mechanisms were designed and controlled by the State in every single detail. Thus, it seems reasonable for them to have served a dominant ideology that praised "non-manual" and "intellectual" work, and a labour market system where the only secure and well- paid job was that of the "white-collar" , public-sector employee. Job insecurity, exploitation by employers, low wages, bad conditions of work, are all societal factors that pushed--and continue to push--pupils toward the more advantageous public sector. The fact that the public sector has been the biggest employer of graduate labour in the past has had a great impact on the prestige of general secondary or university education, and the "inferiority" of technical-vocational education.

Thus, whereas in the past, winning a place in the upper- secondary school had been considered a success, in the last decade there has been a strong public pressure for "freer" access to the universities and the other institutes of higher education. But since places in higher education--given its "free-of-charge" character, and at the same time the restraints imposed by the State budgets--were limited, Greece has experienced a situation where the "demand" exceeded the "supply". This imbalance had--and still has--to be controlled by the system of the National Examinations.

Various studies have shown that the allocation of university places is very inequitable and favours high income groups, or groups with high social status. Even when the research findings claim that performance in the National Examinations--and subsequently success in entering the university--is not directly affected by the "family socioeconomic background", there is, nevertheless, always an indirect influence through various other factors.(Polydorides, 1995 a,b and 1996) Factors such as "curriculum track" or "attendance of private cramming institutes" underscore the influence that the family exercises on the choices made, on the one hand, and on the resources used for ensuring eventual success, on the other. The greater ability to finance preparatory classes and enter selective private schools, results in the finding that "...sons and daughters of managers, executives and professionals are four times as likely to enter the university on their first trial relative to the offsprings of manual workers". (Papas and Psacharopoulos, 1987, p.494)

Notes

1. The term is rather simplistic, and it is used purely for a minimum level of descriptive precision.
2. We should not forget her role as a founder member of the E.C., the EURATOM and the ECSC in the 1950s.
3. Indeed, the country's annual industrial production is among the highest in Europe, despite the wide disparities between North and South, and its long tradition of political instability.
4. Here we should note that all the following governments used the distribution of textbooks for strengthening control over knowledge.
5. The proportion of those occupied in the primary sector has diminished dramatically during the last 35 years. From about 54% in 1961, it fell to 29% in 1985, and 21% in 1994. At the same time, the figures for the secondary sector were 19%, 27% and 24%, and for the tertiary sector 27%, 44% and 55.5%, respectively.

References

Archer, M. (1979). Social Origins of Educational Systems. London:Sage.

Ball, S. (1990). Politics and Policy Making in Education: explorations in policy sociology. London: Routledge & Kegan Paul.

Bernstein, B. (1973). Class, codes and control, Vol.1. London: Routledge & Kegan Paul.

Bernstein, B. (1975). Class, codes and control, Vol.3. London: Routledge & Kegan Paul.

Bourdieu, P. (1964). Les heritiers: les etudiants et la culture. Paris: Minuit.

Bourdieu, P. & Passeron, J.C. (1976). Reproduction in Education, Society and Culture. London: Sage.

Broadfoot, P. (1979). Assessment, schools and society. London, Methuen.

Brock, C. & Tulasiewicz, W. (1994). Education in a single Europe. London: Routledge & Kegan Paul.

Carnoy, M. & Levin, H. (1976). The limits of educational reform. New York, Mc Kay.

Christie, T. & Forrest, G.M. (eds.) (1981). Defining public examination standards. Basington: McMillan Education.

DES (1988) The National Curriculum. London:HMSO.

Dimaras, A. (1973). The reform that never happened Vol.1. [in Greek] Athens: Hermes.

Dimaras, A. (1974). The reform that never happened Vol.2. [in Greek] Athens: Hermes.

Drettakis, M. (1974). The professional orientation in the three-year gymnasium. The Citizen [in Greek].

Elvin, L. (ed.) (1981). The Educational Systems in the European Community: a guide. Windsor: NFER-NELSON.

Gipps, C. AND Murphy, P. (1994). A fair test? Assessment, Achievement and Equity. Buckingham: Open University Press.

Greek Ministry of Education (1994). Statistics for Higher Education. Academic Year 1993-94. [in Greek] Athens: Statistical Service of the Ministry of Education.

Greek Ministry of Education (1996). Minimum scores for entry into various Higher Education Departments. [in Greek] Athens, Statistical Service of the Ministry of Education.

Halsey, A.H. (1977) Towards Meritocracy? The Case of Britain. In: J. Karabel & A.H. Halsey (Eds). Power and ideology in education. New York: OUP.

Halsey, A.H., Heath, A.F. & Ridge, J. (1980). Origins and destinations: family, class and education in modern Britain. Oxford, Clarendon Press.

Kalamatianou, A.G., Karmas, C.A. & Lianos, T.P.. (1988). Technical Higher Education in Greece. European Journal of Education, 23(3), pp.271-279.

Kassimati, K.(1991). A Survey on the Social Characteristics of Labour. [in Greek] Athens: EKKE.

Kassotakis, M. (1992). The school and career orientation of students in the integrated multilateral lyceum: a critical evaluation. Education and Occupation, 3(2-3), pp. 102-127 [in Greek].

Katsikas, C. & Kavadias, G. (1994). Inequality in Greek Education. [in Greek] Athens: Gutenberg.

Kelly, A.V. (1990). The National Curriculum: A Critical Review. London: Paul Chapman.

Kerckhoff, A. & Trott, J.M. (1993). Educational Attainment in a

Changing Educational System: The Case of England and Wales. In: Y. Shavit & H.P. - Blossfeld (Eds.). o.p.

Kokkos, A. (1987). The social role of technical lyceum: educational-professional aspirations and direction of pupils. Unpublished Research Project. [in Greek] Athens: Ministry of Industry, Energy and Technology.

Kostaki, A. (1992). How much "integrated" is the multilateral lyceum?. Education and Occupation, 3(2-3), pp.140-148 [in Greek].

Kyprianos, P.(1996). Social Representations of the University Diploma. Synchrona Themata 19 (60,61), pp.239- 246 [in Greek].

Mallinson, V. (1980). The Western European idea in education. Oxford: Pergamon Press.

McLean, M. (1991). Education in France. In Holmes, B. Equality and Freedom in Education. London: Routledge & Kegan Paul.

National Statistical Service of Greece (NSSG) (1995). Labour Force Survey. [in Greek] Athens, NSSG.

Nikta, A. (1991). Reform of Greek Secondary Education from 1974 to 1989. Unpublished Ph.D. Thesis. Manchester: University of Manchester.

Noutsos, C. (1979). Secondary School Curriculum and Social Control. [in Greek] Athens: Themelio.

Nuttall, D. (1987). The Validity of Assessment. European Journal of Psychology of Education, 11, pp.109- 118.

OECD (1980). Educational policy planning: educational reform policies in Greece. Paris: OECD.

Papas, G. (1993). The Determinants of Educational Achievement in Greece. Studies in Educational Evaluation,17 pp.405-418.

PASOK, (1981). Declaration of Governmental Policy: The Contract with The People. [in Greek] Athens: PASOK.

Polydorides, G. (1990). The Examinations for the Higher Education. EKE, 76, pp.84-111.

Polydorides, G. (1995a). Educational Policy and Practice: a sociological analysis. [in Greek] Athens: Hellenic Grammata.

Polydorides, G. (1995b). Sociological analysis of Greek Education: the University-entry Examinations, vol. 1. [in Greek] Athens: Gutenberg.

Polydorides, G. (1996). Sociological analysis of Greek Education: the

University-entry Examinations, vol. 2. [in Greek] Athens: Gutenberg.

Psacharopoulos, G., & (1987). The transition from school to the university under restricted access. Higher Education, 16, pp. 481-501.

Shavit, Y. & Blossfeld, H.P. (1993). Persistent Inequality: Changing Educational Attainment in Thirteen Countries. Boulder Co.:Westview Press.

The Economist, 4/5/1996 and 6/4/1996.

UNESCO (various years) Statistical Yearbook. Paris: UNESCO.

Young, M. (Ed.) (1971). Knowledge and control. London: Collier-Macmillan.

About the Author

Dionysius Gouvias Research and Graduate School
Faculty of Education
University of Manchester

Humanities Building
Oxford Road
Manchester M13 9PL

Email: mewxddg1@fs1.ed.man.ac.uk

I graduated (with a honour's degree) from the Department of Sociology of the University of Crete, Greece, in 1993. I did my masters' degree (University of Manchester, Faculty of Education) on "Educational Policy", between 1994 and 1995. I registered as a Ph.D. candidate in the Research and Graduate School of the Faculty of Education, University of Manchester, in September 1995.

During the last five years I have participated as a research assistant in two research projects on the socioeconomic transformation of certain regions in Greece. (These projects were funded by the Universities of Athens and Crete.) I participated in two Educational Conferences, one held in Athens (Greece), and one at the Birmingham University, School of Education, in 1996. My main research interests are inequalities in education (mostly in secondary), curriculum content, the relationship between educational policies and cultural settings, and between school and labour market structures.

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

Greg Camilli
Rutgers University

John Covaleskie
Northern Michigan University

Andrew Coulson
a_coulson@msn.com

Alan Davis
University of Colorado, Denver

Sherman Dorn
University of South Florida

Mark E. Fetler
California Commission on Teacher Credentialing

Richard Garlikov
hmwkhelp@scott.net

Thomas F. Green
Syracuse University

Alison I. Griffith
York University

Arlen Gullickson
Western Michigan University

Ernest R. House
University of Colorado

Aimee Howley
Marshall University

Craig B. Howley
Appalachia Educational Laboratory

William Hunter
University of Calgary

Richard M. Jaeger
University of North Carolina--Greensboro

Daniel Kallós
Umeå University

Benjamin Levin
University of Manitoba

Thomas Mauhs-Pugh
Rocky Mountain College

Dewayne Matthews
Western Interstate Commission for Higher Education

William McInerney
Purdue University

Mary P. McKeown
Arizona Board of Regents

Les McLean
University of Toronto

Susan Bobbitt Nolen
University of Washington

Anne L. Pemberton
apembert@pen.k12.va.us

Hugh G. Petrie
SUNY Buffalo

Richard C. Richardson
Arizona State University

Anthony G. Rud Jr.
Purdue University

Dennis Sayers
University of California at Davis

Jay D. Scribner
University of Texas at Austin

Michael Scriven
scriven@aol.com

Robert E. Stake
University of Illinois--UC

Robert Stonehill
U.S. Department of Education

Robert T. Stout
Arizona State University

archives | abstracts | editors | board | submit | comment | subscribe |

search

EPAA
• enter the archives • browse the abstracts • the editors • the edit board
• submit article • submit commentary • search • subscribe
volume: • 1 • 2 • 3 • 4 • 5 • 6 • 7

This article has been retrieved 941 times since February 5, 1998.

Education Policy Analysis Archives

Volume 6 Number 5 February 5, 1998 ISSN 1068-2341

A peer-reviewed scholarly electronic journal.
Editor: Gene V Glass Glass@ASU.EDU.
College of Education Arizona State University,Tempe AZ 85287-2411

Copyright 1998, the EDUCATION POLICY ANALYSIS ARCHIVES.Permission is hereby granted to copy any article provided that EDUCATION POLICY ANALYSIS ARCHIVES is credited and copies are not sold.

A Remarkable Move of Restructuring: Chinese Higher Education

Fang Zhao

**University of Western Sydney
Australia**

Abstract

In this article, the current remarkable trend of institutional amalgamation and the establishment of cross-institutional consortiums in China are examined. The principal purpose of this study is to explore policy options on issues connected with the trend and the significant implications of the trend for the future development of higher education in China. I discuss the outstanding issues raised in the restructuring, the main factors behind them and proposes policy options to redress the adversities of the trend at the end. The article draws on national data as well as a case study. The research reported here is constructive for comparative and empirical research of similar issues in international perspectives.

Introduction

The dilemma between rapid growth of higher education and increasing financial constraints has led to an increasing emphasis on the need to improve efficiency by better utilisation of resources. Like elsewhere, in China, attempts have been made to optimise educational funds through institutional mergers and cooperation between institutions in sharing resources, with the intention of raising student-staff ratios and cost-effectiveness. Between 1992 and 1995, more than 70 institutions were merged into 28 institutions and over 100 institutions set up cross-institution consortiums. (Zhu.Kaixuan, 1995) This remarkable trend is a focus of this paper. Hopefully, the research reported in this paper is constructive for comparative and empirical research of similar issues in international perspectives.

The World Bank (The World Bank, 1987) maintained in a mission report on Chinese universities that an increase in the student-teacher ratio could significantly reduce unit recurrent cost, given that the ratio in China is much lower than the average ratio in East Asia and the Pacific. Also, it was found after an analysis of data elicited from 136 Chinese universities that there was an underlying relationship between unit recurrent cost and the size of enrollment; and that the larger an institution, the lower unit recurrent cost.

Pennington (1991) held from his experience of Australian amalgamations between universities and colleges of advanced education that many problems and difficulties must be weighed against the benefits which have accrued or may yet accrue on the amalgamations. He pointed out some major problems, such as the risks of loss of independence and diversity of the amalgamated institutions and of collegial commitment and staff's morale. Karmel (1992) suggested that the benefits of larger institutions were not yet established and held that smaller institutions promoted innovation. Williams (1988) cast doubt on "bigger is better," evidenced by some of the greatest universities in Europe and North America which were smaller than the universities in Sydney, Queensland and Melbourne. Gilbert (1991) indicated that amalgamation contributed to the emergence of a larger, more differentiated, less well resourced university sector than its predecessor.

In China, a recurring theme in the current literature is enthusiasm for boosting consolidation and cooperation of higher education institutions, while few articles deal with difficulties and problems underlying the trend. Li Peng (1995), Chinese Premier, suggested that jointly-running institutions including consolidation and cooperation of institutions may optimise educational resources. Zhu Kaixuan (1995), director of the State Education Commission (SEC), proposed that conditions would be created to promote consolidation of those small institutions with a narrow range of specialities and redundant courses; and that those institutions within close proximity but with different disciplines be encouraged to set up cooperative relations in sharing resources, complementing each other and combining disciplines for mutual viability.

Li Zhengyuan (1995), however, had a different opinion that the current pursuit of larger and more comprehensive institutions failed to produce cost-effectiveness and improve the quality of Chinese higher education but caused a false upgrade of education establishments (two-year colleges upgraded as four-year universities when consolidated with universities, for example), duplication and overlapping of organisations, and contrary to expectations, increased staff members due to redundancy. Wang Wenuyou (1995) conducted a survey over 71 institutions in Beijing and concluded that smaller institutions may not be inefficient, and that the efficiency and effectiveness was determined by appropriateness of size of class, rationalised course offerings and fulfilment of enrollment quota.

The brief review of related literature above shows that the movement towards consolidation and cooperation between institutions has both strengths and deficiencies within international perspectives. What remains to be explored, however, are policy options on issues connected with the movement, and significant implications of the movement for future development of higher education in China. This is the principal purpose of this study. The study is based upon evidence in the literature, theory grounded in international debates and a case study.

Higher Education Structure: Post-1977

It is necessary to briefly overview the development of Chinese higher education in historical perspective before discussing its current trends and issues. In terms of the expansion of higher education, the most remarkable changes occurred following the 1978 economic reform in China. The changes in higher education structure can be divided into two stages as shown in Table I below. At the first stage between 1978 and 1985, the changes were represented by rapid growth in the number of higher education institutions and enrollments. The second stage, post-1985, was characterised by continuously rapid increase in enrollments but relatively stable viability of institutions without any increase in the total of institutions in 1986 and 1995.

Table I
Development in Institutions and enrollments: 1977-1995

Year	No. of institutions	Total enrollments* (million)	Annual increase (thousand) X_i
1995	1,054	3.05	120
1994	1,080	2.93	290
1993	1,065	2.64	360
1992	1,053	2.28	150
1991	1,064	2.13	-30
1990	1,075	2.16	-20
1989	1,075	2.18	0
1988	1,075	2.18	100
1987	1,063	2.08	90
1986	1,054	1.99	200
1985	1,016	1.79	340
1984	902	1.45	140
1983	805	1.31	130
1982	715	1.18	-120
1981	704	1.30	130
1980	675	1.17	130
1979	633	1.04	173
1978	598	0.86	242
1977	404	0.63	-
			$X = 135$
			$s = 123$

Sources: Ministry of Education and the State Education Commission (1984, 1991); The State Statistic Bureau (SSB) (1992-1996)

*Note: Including all undergraduate and graduate students on campus.

Table I indicates that the average annual growth (arithmetical mean) in enrollments between 1978 and 1995 is 135,000 but the growth is extremely uneven with a large standard deviation of 123,000. The fluctuations included negative growth in 1982 (when students in the first major rise intake in 1978 graduated) and in 1989-1991 (due to three consecutive years' economic entrenchment). Table I also shows that at the first stage of seven years (1978-1985), 612 institutic over a half of the total institutions formed within 47 years since 1949, emerged, leaving no increa in numbers of institutions during the following ten years. The rapid emergence of 612 institution was mostly through upgrading former secondary colleges and polytechnics. The dramatic and rap growth in institutions was intended to accommodate an unprecedented expansion of enrollments without much consideration of the actual capabilities of those newly upgraded institutions. Most them were relatively small in size of enrollments as shown in Table II and not well supported in t human and financial resources as disclosed in the Chinese press (Ribao, 1985).

It was reported that in 1986, 90 percent higher education institutions were below the standar required for an education institution set by the State Council in that year, in terms of staff quality, teaching and research facilities and equipment, student accommodation and libraries. It was belie that it was the devolution of accreditation of two-year colleges and polytechnics to local governments that contributed to the rapid growth in institutions with poor quality before 1987 (Zhongguo Jiaoyubao, 1991).

Table II
Frequency Distribution of Institutions by Size of Enrollments: 1978-1990

Year	Total institutions	300 & below (%)	301-500 (%)	501-1000 (%)	1001-1500 (%)	1501-2000 (%)	2001-3000 (%)	3001-4000 (%)	4001-5000 (%)	50 & ov (%
1990	1,075	5.4	7.3	20.7	21.3	13.6	16.0	6.51	2.7	6.
1988	1,075	5.6	8.0	21.9	20.7	13.7	14.0	6.2	3.3	6.
1986	1,054	9.4	9.0	22.8	19.3	12.5	12.6	5.2	3.0	6.
1984	902	12.3	10.3	28.4	20.2	9.2	10.8	4.8	2.3	4.
1982	715	8.4	9.7	31.0	17.9	11.2	10.9	5.5	3.6	1.
1980	675	11.7	8.0	30.2	17.0	11.4	11.1	4.7	3.3	2.
1978	598	16.6	12.5	29.1	16.4	10.4	8.7	4.7	1.0	0.

Sources: Adapted from the State Education Commission (SED) and the Ministry of Education, (1984-1991)

Table II provides detailed statistical information about changes in size of enrollments or size of institutions between 1978 and 1990 (no national and official data available after 1990). In 1978, institutions with 501 to 1,000 students accounted for 29.1 percent of all institutions, the modal percentage, followed by institutions with fewer than 300 students; whereas in 1990, the biggest percentage of all institutions, 21.3 percent, accrued for institutions with a range of 1,001 to 1,500 students, followed by 20.7 percent of institutions with 501 to 1,000 students.

Table III

Cumulative Frequency Distribution of Institutions by Size of Enrollments: 1978-1990

Year	1 or more %	301 or more %	501 or more %	1001 or more %	1501 or more %	2001 or more %	3001 or more %	4001 or more %	5001 or more %	Mdn. enrollment
1990	100	94.6	87.4	66.7	45.4	31.8	15.8	9.3	6.6	1393
1986	100	90.6	81.6	58.8	39.5	27	14.4	9.2	6.2	1229
1982	100	91.6	81.9	50.9	33	21.8	10.9	5.4	1.8	1026
1978	100	83.4	71	41.9	25.5	15.1	6.4	1.7	0.7	862

Sources: Adapted from the State Education Commission (SED) and the Ministry of Education, (1984-1991)

Table III further displays a clear trend of development towards bigger institutions. In 1978, only 0.7 percent of institutions had an enrollment of more than 5,000 students but in 1990, the number of such institutions rose to 6.6 percent. In 1978, 83.4 percent of institutions had more than 300 students and the number rose to 94.6 percent in 1990. Also, the median enrollment was only 862 students and it rose to 1393 in 1990, an increase of 61.6 percent. Despite the increase, only about 15 percent of institutions had more than 3,000 students in 1990 and much fewer prior to 1986.

In 1986, to restrain the extremely fast growth in institutions and improve the quality of higher education, the State Council circulated the Provisional Regulations on Establishing Higher Education Institutions, and revoked the accreditation of higher education by local governments. In 1988, the State Education Commission issued another policy paper to reinforce the quality standard on higher education institutions set by the State Council in 1986 (Zhongguo Jiaoyubao, 1991, October 8, p. 1).

As shown, since 1986, the emphasis on the expansion of higher education began to be shifted from setting up new institutions to adjustment of the structure of existing institutions. Confronting serious tensions raised in the first seven years of expansion, such as growth versus quality and expansion versus cost-effectiveness, the central government also sent a clear message to the higher education sector that no encouragement would be made to build new institutions in the next five years, and that expansion of enrollments was to be achieved through tapping the existing resources and extending the existing institutions (Li, Peng, 1986). Under the guideline of this policy proposal, the exuberant growth of institutions was eased and a trend towards larger institutions began to take shape as shown in Tables I, II and III.

A Recent Trend and Corresponding Issues

A brief overview of the Chinese higher education structure above illustrates that a large gap exists between the rapid growth in participation in higher education, (that is, the national enrollments) and the enrollment capacities of individual institutions which had only limited expansion. The national enrollments increased by 148 percent from 0.86 million in 1978 to 2.16 million in 1990, but the median enrollment of individual institutions rose only 61.6 percent during the same period, as shown in the above tables. Additional enrollments had to be accommodated through the building of new institutions, a costly strategy compared with the expansion of existing institutions. Hence there was an urgent need to enlarge the enrollment size of institutions so as to accommodate the rapid growth in participation in higher education. As the total government revenue as a ratio of GNP was continuously declining from 32.2 percent in 1978 to 21.8 percent in 1985 and to 17.2 percent in 1992 (Zhongguo Jiaoyubao, 1994, October 6, p.1), it was getting harder for the government to afford building new institutions than to enhance the capacity of existing institutions. The financial constraint was a major driving factor for a shift towards institutional consolidation and cooperation with the intention of achieving cost-effectiveness and optimisation of resources.

As early as 1986, the first cross-institution consortium was set up in Beijing which has the largest number of institutions as a municipality in China. The consortium was composed of eight higher education institutions with a total enrollment of 47,000 students and fixed assets of 0.6 billion RMB. The eight institutions set up close cooperative relations in a number of areas, including open access to laboratories, libraries and lecture, exchanging academic staff and teaching materials, cooperation in research and joint-training staff (Zhongguo Jiaoyubao, 1987, July 7, p.1).

However, the development of such cross-institution consortiums was very slow and few and far between, and there was no official report on institutional consolidations before 1992. In late 1992 and early 1993, the Central Government proposed a new round of reform in higher education by concentrating on higher education management. As a part of the reform, consolidation and cross-institution cooperation were highly recommended by the government as a means of optimization of resources (Li, 1995; Zhongguo Jiaoyubao, 1995, July 12, p.1; Zhu, 1995). The most dynamic development of such structural changes occurred in 1995 and prevailed in almost every province in China. In major cities such as Beijing, Shanghai and Guangzhou, where there were more institutions than in other areas, institutional mergers and cross-institutional cooperation were growing more vigorously. In terms of the latest statistical information, more than 70 institutions got involved in institutional mergers, among which 42 institutions consolidated in 1995; and about 100 institutions joined in cross-institution consortiums (Zhongguo Jiaoyubao, 1995, November 24, pp.1-2). In the writer's view, the waves of such consolidation and cooperation will shake up the entire structure of Chinese higher education over the next few years. The significance of restructuring cannot be too great for the future viability of Chinese higher education.

There exists a common belief that institutional consolidation helps achieve cost-effectiveness and optimization of the insufficient resources supplied to higher education, through raising student-teacher ratios, reducing waste and redundancy, and sharing resources (Clark and Neave, 1992). It was on the basis of this common belief that consolidation and cross-institution cooperation were initiated and developed in China. As the trend of consolidation and cross-institution cooperation started not long ago and is still in progress, it is too early to locate much evidence of the actual outcomes of the structural changes. The following discussion is based upon both potential and realities.

In 1995, Shanghai boasted a total of 45 higher education institutions with about 140,000 students on campus. The average enrollment for each institution was 3120 students. However, there were 23 institutions whose enrollments were below 2000 students and 11 institutions with fewer than 1000 students. In the light of a government's plan for restructuring higher education in Shanghai, the 45 institutions will be consolidated into 30 institutions with an average enrollment of over 4680 students. The capacity of enrollment of each institution will increase by 50 percent (Zhongguo Jiaoyubao, 1995, December 4, p.1). It is evident that consolidation is likely to enlarge the institutional capacity of enrollment. But the capacity is also determined by other important factors such as popularity of course offerings, quality and morale of staff, teaching and research facilities, student services, etc. Besides, cost-effectiveness is achieved through increasing student-teacher ratios and removing redundancy. The above mentioned plan did not deal with this sensitive issue, that is, how much redundancy would be cut to achieve efficiency, as student-teacher ratios were very low with around a 7.3:1 ratio nationally by the end of 1995 (SEC, 1996).

There was a news report about increasing student-teacher ratios from 5.5:1 in 1991 to 8.1:1 in 1994 through restructuring higher education institutions administered by the Ministry of Internal Trade in China (Zhongguo Jiaoyubao, 1994, August 25, p.1). This is one of a few successful cases reported as having achieved a relatively high ratio of students and teachers through consolidation.

In a World Bank mission report on Chinese universities in the late 1980s, it was found that substantial economies of scale existed in university operations in China. By a statistical analysis of data submitted from 136 Chinese universities, the mission reported that there was a generally declining average recurrent cost for institutions of larger size in the 136 universities. Based upon the sample of the 136 institutions, the mission also made a simulated measure of the effect of increases in enrollment and in student-teacher ratios on recurrent costs in six different kinds of institutions with a simulated enrollment range from 500 to 15,000. The results suggested that there were significant savings in terms of lower average unit cost up to a level of about 8,000 to 10,000 students and further expansion would lead to less substantial reductions in unit costs. The results also displayed that much higher savings would be produced if student-teacher ratios of 8:1 by 1990 and 12:1 later could be achieved in terms of an approximate target set by the SEC, closer to an international average. So the mission recommended that smaller institutions operating in close proximity should consider the possibility of consolidation under a single administration (World Bank, 1987).

As shown above, the realised and potential benefits of larger institutions imply greater opportunities to expand higher education by tapping existing resources without injecting additional funding. The World Bank survey also provided evidence in favour of larger institutions and raised student-teacher ratios to produce substantial savings. However, the above reports including that of the World Bank failed to look at or estimate possible difficulties and problems institutional consolidation may raise in practice. A case study of a cross-institutional consortium, a loose federal model of consolidation in fact, is illustrated below to highlight the issues accompanying consolidation.

A Case Study

(This case study drew heavily on an official journalistic report published in Zhongguo Jiaoyubao, 1995, April 20, p.3.)

In early 1994, five higher education institutions in Beijing founded a cross-institution consortium called "Eastern University City." The five institutions were Beijing Chinese Medicine University, Beijing Chemical Industry University, Foreign Trade and Business University, Beijing Fashion Design Institute and China Finance Institute. The consortium had one central governing body as a coordination and supervision commission to manage overall business of the consortium. Under the central governing body, there were six sub-committees in charge of academic and administrative affairs, institutional industry and business, and student and staff services of the consortium. The five member institutions still retained their own full administrations, which were separately funded and governed by five different state ministries.

On the foundation of the consortium, the five member institutions reached an agreement on cooperation in a number of areas. The agreement included setting up common basic courses, exchanging faculty members, combining library and laboratory resources, credit transfer, cooperation on research projects and on trading and transferring research products, sharing research achievements, sharing student and staff services, and jointly building and sharing student and young staff residences, and the like.

One year later after the foundation of the consortium, a survey was conducted over the progress of the consolidation. It was found that very limited cooperative programs had materialized but most of the agreed cooperation was not implemented and some cooperative activities failed to achieve the desired results. Of the materialized cooperative programs, the most successful was the operation of the Eastern University City Credit Union which attracted 30 million RMB from each member institution and other investors in one month after its foundation and seemed to have a prosperous future.

However, difficulties and problems were raised when it came to other cooperative areas. As far as exchanging staff was concerned, few arrangements could be made because the five institutions had the same problems of over-supply of staff for some courses and under-supply of staff for other courses. Also, because it was often too late to make changes in overall staff arrangements when the five institutions submitted to coordinating committees their respective staff arrangement plans, as any changes would cause conflict in lecturing timetables of teachers.

As to sharing library and laboratory resources, it was hard for students and staff to use other member institutions' libraries and laboratories due to some complicated approval procedures. Cooperation on research was also infrequent because it was difficult to obtain joint-research grants from the five government agencies which funded and administered each of the five institutions. When it came to consolidation of staff and student services, no progress was made as a result of lack of profits and fears of loss of jobs on the part of the staff who worked at the services. Also, the plan to build shared student and staff residences for the five member institutions failed, due to financial stringency

The above brief description of the findings illustrates that institutional consolidation and cooperation are complicated processes involving full commitments and great efforts of every participant in the agreed areas - teaching, research and services. Problems raised in the processes of the consolidation of the five institutions can be summarized as:

- lack of a powerful central administration with clearly-defined roles and responsibilities to ensure cooperation plans were enforced;
- lack of materialized support rather than the rhetoric of approval from
- the government agencies which administered the participating institutions to inject sufficient funds into the consolidation;
- pre-occupation with quick economic returns from consolidation;
- fears of losing jobs because of the potential redundancy caused by consolidation;
- concern about losing institutional status; and
- consumption of time in consolidation processes.

These problems may be relevant to other institutional consolidations and cooperation. Failure to realize and solve those problems has led to a loose federal arrangement of the five member institutions, which obviously increased administrative cost with a new central superimposed administration, contrary to the initial objective of achieving greater savings through consolidation. Given that little sharing of resources was realized and few cooperative programs were completed as shown above, the consolidation of the five institutions was in fact unsuccessful and resulted in a nominal rather than an actual consolidation.

This case study, though it may not apply to the entire trend and issues, implies that many difficulties and problems exist such as those of administration, funding and culture in the processes of consolidation and cooperation. The difficulties and problems inherent in the processes need to be fully understood so that positive outcomes can be achieved, and adversities be minimized.

Policy Options

A further analysis of the factors causing the problems in the processes of consolidation and cooperation reveals that there exist some serious weaknesses in the current higher education management system in China. Firstly, consider the problems of administration. Higher education institutions are under the jurisdictions of different government agencies, each of which independently funds and administers a number of institutions. These institutions become in fact subordinates and properties of a certain government agency. If institutional consolidation and cooperation was implemented between institutions under different government agencies, it would be much harder to succeed, as with the case of the five institutions, because greater bureaucracy and more consideration of each government agency's interests would be involved.

To alleviate the adversity, institutional autonomy should be respected by more than lip service from the government agencies. Institutions should also change their tradition of excessive reliance upon the government and keep an "arm's distance" from the government. Fully-fledged institutional autonomy facilitates processes of consolidation and cooperation between institutions under the jurisdiction of different government agencies. This is because institutional freedom facilitates the development of administrative and educational links between institutions without interference of the government agencies concerned. Institutional autonomy when fully implemented can also weaken and modify the current artificial demarcation of external administration over institutions in the system. The artificial demarcation in the current management system has led to considerable waste and managerial inefficiency in terms of duplication and overlapping of course offerings in institutions under different government's agencies and redundancy of bureaucracy.

Secondly, consider problems of funding. The current funding formula in China is still a student-number based one which makes it hard for academics to obtain research funds, let alone get funds for joint research projects between consolidated institutions. This problem may be partially solved through government's earmarked grants and especially through the sale of academic services to industry/business. But funding for basic fundamental research still relies on the government's support, as industry/business will be more interested in research projects with immediate economic returns.

Now that funding difficulty restricts the development of consolidation, the government should provide adequate infrastructure resources for consolidation and cross-institutional cooperation if it believes that such consolidation and cooperation will achieve more economic and social returns in the long run. Also, resource allocation within institutions needs to be improved. Financial responsibility should be delegated to academic departments and research centres to facilitate cross-department collaboration and/or joint research, if this has not been realized. The government should also consider a shift from financing research jointly with teaching to funding it separately to ensure fundamental research and also to provide a springboard for the attraction of supplementary research funds from consumers of research products. The sale of academic services has been evidenced to be a major supplementary source of income for many institutions in China (SEC, 1995), thus alleviating financial stringency of institutions. If a full integration of academic services is fulfilled between institutions, greater savings can be produced by sharing administrative and physical resources, such as having a single administration and joint use of research facilities and equipment.

Thirdly, the cultural problems of consolidation are as important as those of administration and funding discussed above. Efficiency and effectiveness in consolidation can only be achieved where staff fully accept each other, where there is acceptance of common purpose, and where staff are fully committed to a consolidated institution. As reported in the consolidation of the five institutions in Beijing, lack of staff's commitment was partially conducive to the failure of consolidation. When interviewed, some staff expressed their indifference to the consolidation and some revealed their fears of losing status and even jobs (Zhongguo Jiaoyubao, 1995, April 20, p.3).

A possible solution is to promote changes in attitudes and enhance morale of staff. A positive environment for consolidation can be created through convincing arguments and evidence on the benefits for students, staff and institutional management that can accrue from institutional consolidation. Financial incentives and administrative discipline should also be enforced to assist the solution. Furthermore, there should be a collegial and democratic process of decision making on major issues within an institution such as whether to institute consolidation or not. The faculty participation in governance is also an option to enhance staff's spirit or commitment to what they choose to do.

Finally, geographical contiguity, educational links and administrative links are all important factors for implementation of consolidation and cooperation. If consolidation and cooperation is instituted between institutions with these links, it is more likely to succeed in achieving significant cost savings and effectiveness. This can be justified in terms of savings in administrative costs through a single central administration that controls several entities in close proximity and under a common government agency. Consolidation with educational links reduces duplication and overlapping course offerings, as redundancy can be removed through sharing staff between institutions that offer common educational programs. But savings can also be produced by a group of institutions offering complementary courses, because sharing existing resources is much less expensive than setting up a wide range of new courses by each institution. The rationale for consolidation lies in educational and economic benefits of broader institutional profiles, readier access to physical resources and more extensive equipment and facilities. The full benefits will require strong links between institutions in the three aspects: closer location, a common government administration, and common or complementary educational programs.

Notwithstanding the significance of the three factors influencing consolidation, there are other alternative possibilities for institutions to achieve both educational and economic benefits. The writer of this paper suggests the following three models warrant investigation.

Agreement Model

In this model, sharing human and/or capital resources is achieved by agreement between institutions without loss of institutional identity. The rationale for this model is its flexibility, voluntary collaboration and maintenance of diversity. Institutions in this model are free to seek academic partners and facilities from any institutions without artificial barriers of identity. There can be a variety of cooperative models at each level of institutions - cross-institution consortiums, cross-department consortiums, common staff development, joint research programs, reciprocal services, joint use of buildings and sporting fields, etc. Since any cooperation in this model is based upon formal and informal agreements between voluntary partners, some administrative and cultural problems and difficulties raised in consolidation will be avoided. The diversity of institutions is retained as no change of institutional identity is involved in this partnership.

Sponsoring Model

This model represents an arrangement between a large, well-established and well funded institution and a fledgling institution or a poorly funded one. The sponsoring institution provides substantial academic and physical support for the sponsored institution to develop to fully-fledged status. This sponsorship has the outstanding advantage of making full use of existing and potential resources of the sponsoring institution to reduce its potential redundancy. Another advantage is that the academic cultures of both the sponsored and the sponsoring interact in this model. Last, there are no risks of changing or losing institutional status or identity in the cooperation.

Government Model

Government model refers to a government funded and run project. The local government builds up common teaching and research facilities, libraries, student and staff residences and living services that are accessible to all institutions in a local area. The local government funds and runs these facilities through levying an educational tax from local enterprises and residents and through charging institutions a sum of below-market service fees. This model applies to medium and large cities where more institutions are gathered. There are four potential advantages to be seen in adopting this model. First, resources are concentrated in this way rather than scattered in each institution. Second, the concentration of resources promotes the highest quality of teaching and research and alleviates severe shortages of student and staff residences and services, as only concentrated funding can afford to do so in the current financial stringency in China. Third, the accessibility of those facilities helps optimize the use of resources rather than having them under utilized in single institutions. Finally, a single government management of those facilities reduces disputes that may occur in using the facilities between institutions as the local government is the only supplier and coordinator of the facilities and institutions are all customers.

Compared with consolidation, the greatest common benefit of the three models comes from no additional administrative costs involved in these institutional links. In addition, institutional identity and diversity are retained, a matter of great concern in relation to consolidation. In view of the objectives of consolidation and cooperation, any appropriate arrangement of links between institutions is highly recommended so long as efficiency and effectiveness in using resources are achieved to its greatest extent.

Conclusion

The purpose of this article was to explore policy options connected with the current trend of institutional consolidation and cooperation and the significance of this trend for the future development of higher education in China. Through discussing the factors behind the trend from an historic perspective, the paper identifies and analyses critically the current trend and its corresponding issues. To address these issues, this paper provided a series of policy options to be implemented or investigated. The study of this paper concluded that the trend to consolidation and cooperation will develop further as a result of pressures from the government and the economy. The development of the trend implies that higher education institutions will grow larger with more capacity for enrollments, broader educational profiles and more concentration of resources with potential cost savings. On the other hand, the trend is also likely to generate a series of acute consequences shown below:

- more managerial and centralized processes of administration as larger and sophisticated institutions require more powerful central control;
- more pressures for partners to combine against their wishes;
- more extensive academic drift through colleges consolidating with universities;
- narrower range of teaching and research activities to achieve economies of scales;
- lower academic quality and standards due to highly increased workloads of staff and normative upgrade of status through consolidation;
- less diversity in the nature of courses and approaches to course provision; and
- more industrial disputes in view of varying wages and different standards for staff's promotion between institutions.

For policy makers and university managers in other countries, the Chinese experience and the discussion of both potential benefits and adversities of the trend is worthy of consideration for improving current policy and practice relating to institutional mergers and consortiums.

References

Clark, B.R. and Neave, G.R. (Eds.). (1992). The encyclopedia of higher education. (Vol. 1-2). Oxford: Pergamon Press

Gilbert, A. (1991) Current issues and future developments in higher education. Journal of Tertiary
Education Administration, 13 (2)

Karmel, P. (1992) The Australian universities into the 21st century. Australian Quarterly, 64 (1)

Li, Peng(1986). Keep reforming educational system and give more materialised support to schools.

Li Peng (1995). Developing higher education through reforms. Zhongguo Jiaoyubao (China
Education Daily), 1995, July 12 p1

Li, Zhengyuan. (1995, November 9). Problems in institutional links should be noted. Zhongguo
Jiaoyubao (China Education Daily), p3.

MOE & SEC (1984, 1991). Achievements of education in China (1949-1990). Beijing: Author

Pennington, D. (1991) Amalgamations in higher education in Australia: issues in Australian higher education. Canberra: The Australian Vice-Chancellors' Committee

Ribao, Renmin (People's Daily) 1985, January 20 p5; 1988, March 31 p3; 1988; March 10 p3

Ribao, Renmin (People's Daily,), 1986, December 3 p3

SEC (1995). China's education at a glance: essential statistics for 1994. Beijing: Author

SEC (1996).The ninth five-year plan for educational development and the long range development program toward the year 2010. Beijing: Author

SSB (1993-1996). Statistical yearbook: China. Beijing: Author

Wang, Wenyou, (1995, November 23). It is not advisable to set up too many specialities. Zhongguo Jiaoyubao (China Education Daily), p2

Williams, B. (1988) The 1988 white paper on higher education. Australian Universities Review, 2

World Bank. (1987) China: management and finance of higher education. Washington D.C.: Author

Zhongguo Jiaoyubao (China Education Daily), 1987, July 7 p1

Zhongguo Jiaoyubao (China Education Daily), 1991, April 16 p2

Zhongguo Jiaoyubao (China Education Daily), 1991, October 8 p1.

Zhongguo Jiaoyubao (China Education Daily), 1994, August 25 p1

Zhongguo Jiaoyubao (China Education Daily), 1994, October 6 p1

Zhongguo Jiaoyubao (China Education Daily), 1995, April 20 p3

Zhongguo Jiaoyubao (China Education Daily), 1995, November 24 pp1-2

Zhongguo Jiaoyubao China Education Daily), 1995, December 4 p1

Zhu. Kaixuan, (1995, November 24). Actively and intensively promoting reforms in higher education management system. Zhongguo Jiaoyubao (China Education Daily), p1.

About the Author

Fang Zhao
Faculty of Education
University of Western Sydney
P.O.Box 10
Kingswood, 2145 NSW Australia.
Postal address: 28/74 Hawkesbury Road, Westmead, 2145 NSW Australia

Email address: n9502277@scholar.nepean.uws.edu.au

Ms Fang Zhao was an Associate Professor in one of key universities in China before she started her research in Australia.She was awarded Master of Education degree by an Australian university for her research in higher education and economic development. She is at the final stage of her doctoral research in the governance and funding of higher education.

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

John Covaleskie
Northern Michigan University

Alan Davis
University of Colorado, Denver

Mark E. Fetler
California Commission on Teacher Credentialing

Thomas F. Green
Syracuse University

Arlen Gullickson
Western Michigan University

Aimee Howley
Marshall University

William Hunter
University of Calgary

Daniel Kallós
Umeå University

Thomas Mauhs-Pugh
Rocky Mountain College

William McInerney
Purdue University

Les McLean
University of Toronto

Anne L. Pemberton
apembert@pen.k12.va.us

Richard C. Richardson
Arizona State University

Dennis Sayers
University of California at Davis

Michael Scriven
scriven@aol.com

Robert Stonehill
U.S. Department of Education

Greg Camilli
Rutgers University

Andrew Coulson
a_coulson@msn.com

Sherman Dorn
University of South Florida

Richard Garlikov
hmwkhelp@scott.net

Alison I. Griffith
York University

Ernest R. House
University of Colorado

Craig B. Howley
Appalachia Educational Laboratory

Richard M. Jaeger
University of North Carolina--Greensboro

Benjamin Levin
University of Manitoba

Dewayne Matthews
Western Interstate Commission for Higher Education

Mary P. McKeown
Arizona Board of Regents

Susan Bobbitt Nolen
University of Washington

Hugh G. Petrie
SUNY Buffalo

Anthony G. Rud Jr.
Purdue University

Jay D. Scribner
University of Texas at Austin

Robert E. Stake
University of Illinois--UC

Robert T. Stout
Arizona State University

---

archives | abstracts | editors | board | submit | comment | subscribe | search

• enter the archives • browse the abstracts • the editors • the edit board
• submit article • submit commentary • search • subscribe
volume: • 1 • 2 • 3 • 4 • 5 • 6 • 7

This article has been retrieved 2637 times since March 1, 1998.

Education Policy Analysis Archives

Volume 6 Number 6 **March 1, 1998** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal.
Editor: Gene V Glass Glass@ASU.EDU.
College of Education Arizona State University,Tempe AZ 85287-2411
Copyright 1998, the EDUCATION POLICY ANALYSIS ARCHIVES.Permission is hereby granted to copy any article provided that EDUCATION POLICY ANALYSIS ARCHIVES is credited and copies are not sold.

School Improvement Policy: Have Administrative Functions of Principals Changed in Schools Where Site-Based Management is Practiced?

C. Kenneth Tanner
The University of Georgia

Cheryl D. Stone
M. G. Barksdale Elementary School
Conyers, GA

Abstract

Have administrative functions of principals changed in schools practicing site-based management (SBM) with shared governance? To deal with this issue we employed the Delphi technique and a panel of 24 experts from 14 states. The experts, which included educational specialists, researchers, writers, and elementary school principals, agreed that the implementation of SBM dramatically influences the roles of the principal in management/administration and leadership. Data revealed that the elementary principal's leadership role requires specialized skills to support shared governance, making it necessary to form professional development programs that adapt to innovations evolving from the implementation of SBM.

Introduction

Americans have begun rethinking and redesigning the most fundamental aspects of the way we run our schools--a process known as "restructuring" or "systematic reform" (Fiske, 1995). One of the most widely used approaches to encourage school improvement through this reform effort is site-based management (SBM). Ideally, SBM policy moves control and decision making from the central office to the local building level.

SBM with shared governance represents a major change in the process used to resolve problems. Ideally, instead of problems being resolved from a central location by a staff not directly involved, the local school community settles dilemmas (Caldwell & Wood, 1992). Moving decision-making authority to the building level affords parents, teachers, and students the opportunity to have an active voice in decisions made at the school level. We are, in effect, "creating ownership for those responsible for carrying out decisions by involving them directly in the decision-making process -- and by trusting their abilities and judgments" (Harrison, Killion, & Mitchell, 1989, p. 55). As a result, increased autonomy of the school staff to make decisions at its facility is the expectation. With the expectation of change in the principalship and the demand for the principal to maintain a high level of performance, Wohlstetter and Odden (1992) assert that it is necessary to establish a clear definition of the role of principals.

"Although site-based management appears in many guises, at its core is the idea of participatory decision-making at the school site" (David, 1996, p. 6). Inherent in SBM is the expectation that the role of the principal will change. In particular, those people nearest to the problems, issues, and situations are included in the decision-making process (Goodman, 1994). Critical to the effectiveness of restructuring is the encouragement of teachers to participate in problem solving and decision making (Thurston, Clift, & Schacht, 1993). This job is the major responsibility of the principal, and the key individual identified as instrumental in determining the success of schools is the principal (Krug, 1993).

Background for the Study

In analyzing the emerging role of the principal in the 1990s, Hallinger, Bickman, and David, (1990) concluded that the leadership of the principal is an intricate, context-dependent set of behaviors and processes. The larger, prevailing context is change, and change in the role of the principal is essential to any reform that is to be both quick and lasting (Carlin, 1992). Daniels (1990) in discussing his leadership role in SBM stated that

> While the principal ultimately remains accountable for what happens at the school level, the school's steering committee plays an active role in nearly all decisions made . . . I gave up veto power in an effort to gain the trust and commitment of the staff. (p. 23)

Findings reported by Wohlstetter and Briggs (1994) from their study of 25 elementary and middle schools in 11 school districts in the U.S., Canada, and Australia underscore the status of the role of the principal changing from being the primary decision maker to one of empowering others. Further, Wohlstetter and Briggs found that the most effective principals involved in SBM made available four critical resources to teachers and community members: power, knowledge and skills training, information, and rewards. As a result of the investigation by Aronstein and DeBenedictis (1991), four basic processes of what administrators do when they manage SBM schools surfaced: Principals are to work collaboratively with staff members to analyze problems, set need priorities, resolve issues, and use group dynamics skills.

In the early, developmental stages of SBM, Lindelow (1981) suggested that in the implementation of school-based management, the jobs and functions of the principal would change from those of middle manager for the district to the leader of the school. Over a decade later Wohlstetter (1995) acknowledged that

> The schools where SBM worked had principals who played a key role in dispersing power, in promoting a school wide commitment to learning, in expecting all teachers to participate in the work of the school, in collecting information about student learning, and in distributing rewards. (p. 24)

Principals have moved from middle managers to leaders at the school site. Principals in Goldman's (1991) study indicated that their primary role in SBM became one of supporting people and being the advocate for their work. Talking to others and coaching and looking for opportunities to positively interact become the everyday expectations of the principal's job.

Even though research provides insight into the emerging role of the principal in the 1990s, Drury (1993) states: "it appears that the traditional role of the building principal is in a state of transformation, but that the ultimate result remains to be seen" (p. 19). To increase the likelihood that schools carrying out SBM are effective, the necessity to clarify the roles of the principal has surfaced (Gleason, Donohue, & Leader, 1996; Guskey & Peterson, 1996).

Three themes emerged from the literature as basis for this study:

1. The establishment of the administrative roles of the individual who occupies the position of school building principal is a controversial issue that is pervasive in the educational community (Blase, 1987; Stephens, 1987).
2. A new form of leadership is necessary to effectively support the processes involved in the implementation of school-based management at the site level (Doud, 1989; Vann, 1996; Wohlstetter & Briggs, 1994).
3. "The key role change [in SBM] is the principal's shift from top-down manager to a supporter and facilitator who maintains his or

> her leadership responsibilities" (Spilman, 1996, p. 36). "Teacher involvement in certain kinds of decisions can be mutually enhancing: it returns to teachers the power to govern their own professional affairs, and teachers, in turn, empower administrators to make decisions that enhance the organization's goals" (Conway & Calzi, 1996, p. 49).

Purpose of the Study

With the policy trend toward the use of SBM influencing school operations, the purpose of this study was to detect changes in selected administrative functions (leadership, decision making, and management) of the principalship. Another purpose was to discover the components of a job profile for elementary school principals working under SBM with shared governance. To this end, a sample of practitioners and educational researchers participating in various aspects of SBM was polled through the Delphi method.

Research Questions

Based on the aim of this study, the following research questions were generated as a guide:

1. What changes have occurred in the principal's role with respect to management and administration after the implementation of SBM?
2. What changes have occurred in the elementary principal's role with respect to leadership after the implementation of SBM?
3. What are the primary management and administrative tasks of the elementary principal in SBM?
4. What are the primary leadership tasks of the elementary principal in SBM?
5. How does the implementing of SBM policy alter the role of the elementary principal in the decision-making process?

Value of The Study

In the past, the organizational structure within school districts has supported the strategy of exerting control over the operations and personnel at the local school from a central office. One prevalent plan to decentralize the organizational management system is the implementation of SBM. However, documentation of the roles and primary tasks of the elementary school administrator participating in SBM with shared governance has not been completed. Building level administrators working in SBM need basic guidance in planning for professional and personal growth. The results of this study are expected to be of value in training programs, establishment of evaluation guidelines, and identification of leadership skills for educational administrators.

During the transition to school-based management, many principals may be asked to assume responsibilities for which they are unprepared or for which their preparation has become dated.

Therefore, development of the job description and principal selection criteria for principals in SBM schools are crucial. The primary functions of the principal in SBM identified in this study may be beneficial to school systems requiring the performance of specific roles and tasks of principals. As a result, applicants and job criteria may be more effectively matched.

Method

The need to clarify the roles of the principal provided a sound basis to select a method of inquiry involving consensus building. Consequently, the Delphi technique was selected. We assumed that people who do the work should be involved in defining roles of their jobs. "Ultimately, it will be the people who carry out site-based management that determine what it is--and can become" (David, 1996, p. 9).

To reach the goal of clarifying the principal's role, the study focused on discovering the functions most often performed by principals in schools operating under SBM policy. Emphasis was placed on narrowing and refining responses of the selected expert panel to a consensus of opinion (Putnam, Spiegel, & Bruininks, 1995; Tanner & Williams, 1981).

The Delphi Technique

Early Delphi studies originated at The Rand Corporation with Olaf Helmer (1967) and his colleagues. These studies involved a systematic method of eliciting expert opinion on a variety of topics with a focus on scientific and technological forecasting (Sackman, 1974). Putnam, Spiegel, and Bruininks (1995) described Delphi as a process to determine opinions or judgments of a group of people. "Delphi may be characterized as a method for structuring a group communication process so that the process is effective in allowing a group of individuals, as a whole, to deal with a complex problem" (Linstone & Turoff, 1975, p. 3).

Cycle I

Uhl (1983) asserted that in the traditional Delphi study panel members are given the opportunity to provide responses to unstructured questions. The panel members in Cycle I were asked to respond to a query soliciting their perceptions regarding the job of the elementary principal involved in SBM (Table 1). No background information or definitions of SBM were included with the questionnaire to avoid influencing the opinions of the participants.

Table 1
Cycle 1: The impact of SBM on the roles of elementary school administrators

Please respond to the open-ended questions below: List item statements that define the roles of the elementary principal whose school is involved with SBM with shared governance. Include any additional comments for other areas that would provide for a more comprehensive Profile of the Elementary School Principal.

- I. How has your job changed in carrying out site-based management with respect to?
    - Administration
    - Management
    - Leadership
    - Other
- II What are the elementary school principal's primary tasks in?
    - Administration
    - Management
    - Leadership
    - Other

Survey Instruments

In order to guarantee that the Delphi statements reflected the panelists' intent, a semantic analysis was conducted on the written replies. To begin the analysis, two individuals were named coders. They had the responsibility to develop sets of responses similar to those of the expert panel members. During the first step in the semantic content analysis process, each written statement was recorded on an index card. Next, index cards were categorized into sets of responses with each set representing one content idea. The last step consisted of formulating one Delphi item statement to represent each set of responses.

Criteria for Consensus

Criteria for convergence of opinion was established before the study. In determining whether convergence of opinion was reached between cycles, the following criteria were established:

> (1) At least 60% of the responders must be in agreement (Skutsch, & Schofer, 1975).
>
> (2) There is no significant change ($p < .01$) in views between Cycles, indicating that stability has been reached (Linstone & Turoff, 1975).

Panel Selection Criteria and Process

The national panel of experts consisted of two subsets: 12 school principals in elementary schools that had worked in SBM for at least three years, and 12 professionals (authors, researchers, professors, consultants, and administrators) who had attained national, regional, or local recognition as knowledgeable educators in the area of SBM. This second category was identified in the study as specialists.

Efforts were made to eliminate potential researcher bias by devising a nomination process for selecting expert panel members. First, an extensive review of the SBM literature published in 1988-1995 was performed. From this a pool of prominent educators, school districts, and organizations involved in SBM was compiled. Members of this pool were contacted to nominate potential panel members.

Each member of the pool was contacted by telephone and given the opportunity to nominate an expert panel member. Expert panel members were to satisfy one of the following criteria: (a) persons who had written about SBM from field experience or university settings and had been published in a nationally distributed journal within the last five years, (b) individuals whose schools had been identified in a nationally distributed journal because of participation in SBM, (c) investigators who had done studies related to SBM, (d) persons who had conducted training and coordinated programs related to SBM for national, regional, or local organizations, (e) educators who had received recognition in a nationally distributed journal, (f) individuals who had held positions with a national, regional, or local organization or a higher learning institution involved in the implementation, research, teaching, or training in relation to SBM, and (g) principals who had held a position in an elementary school implementing SBM for at least three years. Principals assigned to SBM schools, but who lacked three years' experience as an administrator carrying out SBM, were excluded from this study.

Each nominee was contacted by telephone and asked to participate in the national expert panel for this study. During the telephone discourse, the purpose and significance of the study, the time frame, criteria for expert panel consideration, and the responsibility of participants were explained. Each person was assured anonymity. Special effort was made to have participation of a representative expert in as many different regions of the United States as possible. Calls were stopped when 12 specialists and 12 elementary school principals agreed to be members of the panel of experts. A biographical account of the selected panelists is provided in the Appendix. Letters to confirm each panel member's participation were sent with the Cycle I questionnaire.

Presentation and Analysis of Data

Cycle I

The initial mailing included a cover letter, the questionnaire (Table 1) with detailed directions for its completion, and a stamped return- addressed envelope (N = 24). The phrasing of question one in the Cycle I instrument for principals was different from the same query for specialists. Principals were asked about changes in their job in SBM with respect to administration, management, and leadership in question one. Specialists were asked about changes in the elementary school principalship in SBM with respect to these same three areas.

After four weeks, non-respondents were contacted by telephone to encourage return of the Cycle I instrument. A follow-up postcard was sent to confirm the telephone contact. By the end of September, 22 of the 24 experts had returned their completed instrument. The two non- respondents changed occupational positions, and neither responded.

A total of 513 responses were received, 140 of which addressed change in management and administration. Sixty-four (64) written responses cited changes in the elementary principal's roles with respect to leadership. There were 188 panel statements regarding the primary tasks of principals in the area of management and administration, while 121 statements related to the primary tasks of the principal in the area of leadership. The semantic content analysis conducted on these data resulted in the formulation of 57 Delphi item statements for the Cycle II survey instrument.

Cycles II and III

A "bogus statement" was inserted as item and a distorted group answer was reported for this item. The purpose of the "bogus statement" was to assess the ability of the survey instruments to withstand manipulation by the researchers (Cyphert & Gant, 1971). The survey instrument consisted of 58 Delphi item statements. In Cycle I, a majority of the respondents commented that administrative and management tasks of the elementary principal are too similar and tend to overlap. Panelists suggested that these two categories be combined in succeeding cycles. This suggestion was followed.

An external review panel was utilized to confirm the proper formulation of the Delphi statements prepared for the Cycle II instrument as suggested by Linstone and Turoff (1975). The review panel consisted of ten educators. Four members were teaching in a school implementing SBM. They were asked to review the survey item statements for content validity by making a comparison to the original responses received form the expert panel in Cycle I. The other six reviewers, at another location, were asked to examine the final survey instrument for clarity.

Reviewers were asked to report the length of completion time. This information was included in the cover letter to the expert panel members. Suggestions and comments from both review groups were used in construction of the Cycle II survey instrument.

The Cycle II survey instrument was developed in October 1995 (See Table 2 for the 58 statements). A packet including the survey instrument, cover letter, and return envelope was mailed to the remaining 22 members in the first week of November 1995. A fax number was included for the convenience of panel members who wanted to return the survey electronically.

For identification purposes, each panel member's name was entered at the top of the instrument. Detailed instructions were also included on the first page. Each responder was asked to indicate his or her level of agreement with each statement on the following scale: Strongly Disagree (1), Disagree (2), Agree (3), and Strongly Agree (4).

At the end of November, non-respondents were contacted by telephone and encouraged to return the Cycle II instrument. Postcards were sent to confirm the telephone conversations.

By the first week of December, 21 of the 22 panel members had returned their Cycle II Delphi survey form. The final respondent's survey instrument for Cycle II was not received until the first week in January, which was too late to be included in the Cycle II data tabulation. This panel member was dropped from the study.

In Cycle III, the mode for each Delphi item in Cycle II was reported to the panel. Before providing responses to Cycle III statements, each panel member saw the group mode and his or her response per item to Cycle II. With this information in mind, each person was asked to consider a new response in light of the modal response or state a reason for not changing the Cycle II response.

When the mode for each Delphi item is presented in the findings of this study, it is reported as the most frequently selected numerical scale value. For the "bogus item," number 44, the highest frequency (15) scored by the experts was a scale value of "3" (agree). To find out if the distortion of data by the researcher would be rejected by the panel, the "bogus item" was reported as a scale value of "1" (strongly disagree).

Cycle III was mailed to 21 participants, and 21 surveys were returned. Several statistical procedures were performed on the data obtained from Cycles II and III. The major objective was to determine consensus.

Table 2, reveals the mode, reported by item, and the highest number and percentage of respondents in agreement after Cycle III. Agreement was reached on 51 out of 58 (87.9%) Delphi item statements. A total of 19 experts were in agreement (90.5%) on two items (40 and 52). (For purposes of verification or reanalysis, the entire data set from this study is available for downloading here in either ASCII or Excel Spreadsheet format.

Table 2
Responses to Cycle III (n = 21)

		Experts in Agreement	
Item	**Mode**	**n**	**%**
Changes With Respect to Management/Administration			
1. The principal makes fewer unilateral decisions	4	16	76.2
2. The principal has an ex- panded role in administration	3	13	61.9
3. Time management is more crucial because of the increased responsibility regarding the orchestrating of shared decision- making	4	17	81.0

4. Instead of the principal being singularly responsible for the attainment of the school's goals, all collaborating parties share this responsibility.	4	18	85.7
5. There is an increased responsibility for the principal to build consensus among constituencies.	4	16	76.2
6. The principal delegates more responsibility as a result of having to spend more time involved in a broader array of decisions.	3	14	66.7
7. The principal has more of a commitment to the empowerment of teachers in decision-making.	4	16	76.2
8. The principal has more responsibility in managing decisions at the site level (e.g., Issues the School Leadership Team will resolve).	3	14	66.7
9. There is more need for the principal to expand his/her knowledge base in such areas as group process and inter- personal skills.	4	17	81.0
10. The principal has more responsibility in managing resources.	4	11	52.4
11. The principal has an increased responsibility in managing personnel (e.g., Recruitment of personnel, staffing, defining specific jobs, evaluating personnel performance).	2	12	57.1
12. The responsibility of the principal has increased to function more as a liaison between the community and the school.	3	15	71.4
13. The need has increased for the principal to stay abreast of current research/ educational issues.	4	13	61.9
14. The principal continues to be responsible for the ongoing, day to day work in the school.	4	16	76.2
Changes With Respect to Leadership			
15. The principal has become more of a facilitator of the decision-making process.	3	14	66.7
16. The principal has an increased responsibility to build consensus among all constituencies.	4	15	71.4
17. The responsibility of the principal has increased to cultivate leadership from the ranks of teachers.	4	18	85.7
18. There is an increased need for the principal to have more communication with people on a consistent basis--both oral and written.	3	11	52.4
19. The principal has an increased responsibility to provide teachers with the information needed to reach decisions.	4	14	66.7
20. The nature of site-based management demands that admini- strators develop extensive "people skills."	4	15	71.4
21. The principal has moved away from being the instructional leader at the school to a school manager focused on developing decision-making processes that involve various stake holders.	2	14	66.7

22. The principal must spend increased amounts of time net- working with other schools, professional groups, and community/business groups.	3	14	66.7
Primary Tasks in Management/Administration			
23. Building consensus.	3	16	76.2
24. Staying abreast of the work of the whole school while allowing people to assume responsibility for their part.	4	16	76.2
25. Dispersing information among various school constituencies so that all are informed and have information necessary for making decisions.	3	11	52.4
26. Developing a School Improvement Plan (SIP) through strategic planning.	4	13	61.9
27. Facilitating the involvement of others in school decision- making.	4	17	81.0
28. Coordinating among all the school's constituencies (site, system, community, state, federal, union).	3	15	71.4
29. Carrying out the ideas developed by the group.	3	17	81.0
30. Orchestrating meetings.	3	16	76.2
31. Serving as the manager of people at the site level (e.g., Providing for the recruitment selection, development, evaluation and, if necessary the separation of faculty and staff members who work in the school).	3	16	76.2
32. Maintaining a safe and orderly school environment.	4	13	61.9
33. Creating organizational structure (e.g., Work teams) for school that involves all faculty members in decision- making.	4	16	76.2
34. Facilitating programs by management of resources.	3	15	71.4
35. Recognizing all "SUCCESSES."	4	16	76.2
36. Providing school-wide staff development on a continuous basis.	3	14	66.7
37. Monitoring site activities in terms of what is legal.	3	14	66.7
38. Facilitating research/ data gathering in support of the work of the governance team.	3	15	71.4
39. Managing groups day to day.	3	14	66.7
40. Promoting the vision and the mission of the school.	4	19	90.5
41. Overseeing the budget.	4	12	57.1
42. Overseeing the operation of the school in areas such as building maintenance, safety, transportation, etc.	3	12	57.1
43. Seeing that the SBM Council (school leadership team (SLT), governance team, etc.) elections are held.	3	14	66.7
44. Coordinating the social services provided to families in the community.	3	13	61.9
Primary Tasks in Leadership			

45. Coaching.	3	15	71.4
46. Building consensus.	3	13	61.9
47. Facilitating the involvement of others into decision-making.	4	18	85.7
48. Building a school-wide vision of what can be accomplished.	4	17	81.0
49. Promoting strategic planning for school improvement efforts.	4	17	81.0
50. Providing opportunities for professional growth for all staff.	4	18	85.7
51. Promoting team spirit.	3	12	57.1
52. Keeping the staff informed.	4	19	90.5
53. Communicating with all the school's constituencies.	4	18	85.7
54. Facilitating the change process.	4	18	85.7
55. Organizing meetings.	3	17	81.0
56. Overseeing the operation of the school (budgeting, scheduling, hiring, etc.).	4	16	76.2
57. Carrying out demo- cratically made decisions.	4	14	66.7
58. Helping the School Leadership Team members to build coalitions for the greater good of all students.	4	17	81.0

A statistical comparison between Cycle II and Cycle III is shown Table 3. The variability from the mean for each Delphi statement is shown as well as the change in the standard deviation. In addition to the modes, means and standard deviations were calculated for more in-depth analysis of the convergence of opinion. "The mean and standard deviation, taken together, usually give a good description of the nature of the group being studied" (Borg & Gall, 1983, p. 366).

Means of Cycle II ranged from 2.38 to 3.80. The highest mean score (Mean = 3.80) was reported for item number 40. In item number 40, panel members concurred that a primary task of the principal in SBM is to promote the vision and mission of the school. Item number 21 received the lowest mean score (Mean = 2.38) in Cycle II. Panelists did not agree that the principal's role changed from instructional leader to school manager in SBM.

Mean scores in Cycle III ranged from 2.33 to 3.90. The largest means (Mean = 3.90) for Cycle III were recorded for Delphi statements 40 and 52. Experts emphasized, again as in Cycle II, that promoting the vision and mission of the school (item number 40) is a primary task of the principal in SBM. For item number 52, experts were in agreement that a primary task of the principal is to keep the staff informed. Item number 21 received the smallest mean score (Mean = 2.33) for Cycle III. In Cycle III, more of the participants' opinions converged to the group response of disagreement with the Delphi statement (number 21), which indicated that the principal's role has changed from instructional leader to school manager.

Table 3
Report of Means and Standard Deviations for Cycles II and III and the Difference in Standard Deviation by Item

	Cycle II (N=22)		Cycle III (N=21)		Cycle II to Cycle III Change in SD
Item	**Mean**	**SD**	**Mean**	**SD**	
1	3.76	.436	3.76	.436	0.000
2	3.15	.745	3.10	.641	-0.104
3	3.62	.669	3.71	.644	-0.025
4	3.71	.561	3.81	.512	-0.049
5	3.52	.602	3.76	.436	-0.166
6	3.25	.716	3.15	.587	-0.129
7	3.62	.498	3.76	.436	-0.062
8	3.10	.641	3.25	.550	-0.091
9	3.71	.561	3.76	.539	-0.022
10	3.25	.786	3.40	.754	-0.032
11	2.76	.768	2.57	.746	-0.022
12	2.76	.539	2.81	.512	-0.027
13	3.33	.730	3.52	.680	-0.050
14	3.57	.746	3.71	.561	-0.185
15	3.43	.507	3.33	.483	-0.024
16	3.52	.512	3.71	463	-0.049
17	3.67	.577	3.81	.512	-0.065
18	3.43	.598	3.38	.590	-0.008
19	3.43	.676	3.57	.676	0.000
20	3.71	.463	3.71	.463	0.000
21	2.38	.865	2.33	.730	-0.135
22	2.81	.814	2.81	.680	-0.134
23	3.10	.625	3.10	.625	0.000
24	3.48	.750	3.67	.730	-0.020
25	3.33	.730	3.33	.730	0.000
26	3.10	.995	3.38	.921	-0.074
27	3.43	.746	3.71	.717	-0.029
28	3.10	.831	2.95	.669	-0.162
29	3.14	.793	2.95	.590	-0.203
30	2.81	.750	2.81	.602	-0.148
31	3.05	.740	3.00	.632	-0.108
32	3.48	.602	3.57	.598	-0.004

33	3.48	.602	3.71	.561	-0.041
34	3.30	.571	3.20	.523	-0.048
35	3.62	.590	3.71	.561	-0.029
36	3.38	.590	3.24	.539	-0.051
37	3.43	.507	3.33	.483	-0.024
38	3.19	.680	3.10	.539	-0.141
39	3.00	.775	2.86	.573	-0.202
40	3.80	.410	3.90	.308	-0.102
41	3.33	.730	3.48	.680	-0.050
42	3.29	.717	3.14	.655	-0.062
43	3.10	.944	3.10	.700	-0.244
44	2.90	.768	2.52	.873	0.105
45	3.43	.598	3.29	.463	-0.135
46	3.48	.512	3.38	.498	-0.014
47	3.62	.498	3.86	.359	-0.139
48	3.62	.498	3.81	.402	-0.096
49	3.43	.676	3.81	.402	-0.274
50	3.67	.483	3.86	.359	-0.124
51	3.48	.512	3.43	.507	-0.005
52	3.71	.463	3.90	.301	-0.162
53	3.62	.498	3.86	.359	-0.139
54	3.67	.483	3.86	.359	-0.124
55	3.00	.837	2.86	.573	-0.264
56	3.57	.598	3.71	.561	-0.037
57	3.52	.512	3.67	.483	-0.029
58	3.57	.507	3.81	.402	-0.105

To assess whether stability had occurred for the Delphi items, a t-test was completed for paired samples on each statement for the two subsequent cycles. The t-value statistic was tested at $p < .01$ level of significance. This procedure answered the question, "Did the responses change significantly from Cycle II to Cycle III?" This procedure was used to determine if another cycle of the survey should be conducted.

Seven items failed to meet the criteria for agreement (Items 10, 11, 18, 25, 41, 42, and 51 as shown in Table 4). For the "bogus item," a mode of "3" (agree) was indicated by 13 (61.9%) of the 21 panel members. The "bogus item" was the only Delphi item out of the 58 that showed a decrease in the mean between these cycles and an increase in the standard deviation (0.105). Movement of panel responses from 71.4% to 61.9% consensus indicated that a distorted reporting of the "bogus item" had influenced panel members' responses.

Item number 27, facilitating the involvement of others into school decision-making, received 81.0% group agreement (N = 17). Although agreement was reached on this item, the t-value -2.83 indicated the means for the paired samples were not stable (alpha = .010) between cycles. Consensus was not reached on this item.

Agreement was reached for item number 49. Seventeen experts (81.0%) "strongly agreed" that promoting strategic planning for school improvement is a primary task of the school principal. However, the t statistic indicated that the difference between means from Cycle II to Cycle III was significant at the .01 level. Stability was not achieved [t = - 2.96 (df = 20) (alpha = .008)] and consensus was not reached on this item.

Table 4
Items With No Consensus

		Agreement		Stability		
Item	**Mode**	**na**	**%**	**t**	**df**	**2-Tailed p**
10	4	11	52.4	-1.83	19	.083
11	2	12	57.1	1.71	20	.104
18	3	11	52.4	1.00	20	.329
25	3	11	52.4	.00	20	1.000
27	4	17	81.0	-2.83	20	.010**
41	4	12	57.1	-1.83	20	.083
42	3	12	57.1	1.83	20	.083
49	4	17	81.0	-2.96	20	.008**
51	3	12	57.1	1.00	20	.329

Agreement was defined as at least 60% of the responders (13 or more experts).

**p indicates there was a statistically significant change (p < .01) from Cycle II to Cycle III.

Discussion of the Findings

Research Question One

What changes have occurred in the principal's roles with respect to management and administration after the implementation of SBM? According to the findings in this study, a fundamental change has taken place in the dynamics of the role of the elementary principal. Seven items (shown as item #, statement, and % in agreement) achieving stability and receiving at least 75% agreement reveal this fundamental change:

- 4 : Instead of the principal being singularly responsible for the attainment of the school's goals, all collaborating parties share this responsibility (85.7%)
- 3 : Time management is more crucial because of the increased responsibility regarding orchestrating of shared decision-making (81.0%)

- 9 : There is more need for the principal to expand his/her knowledge base in such areas as group process and interpersonal skills (81.0%)
- 1 : The principal makes fewer unilateral decisions (76.2%)
- 5 : There is an increased responsibility for the principal to build consensus among constituencies (76.2%)
- 7 : The principal has more of a commitment to the empowerment of teachers in decision-making (76.2%)
- 14 : The principal continues to be responsible for the ongoing, day to day work in the school (76.2%)

Other consensus items ranging between 60%-74% agreement were statements 2, 6, 8, 12, and 13. In conjunction with these findings, Black (1996) reports that many principals, the key players in the success or failure of school-based management, are 'paranoid' about their changing roles and responsibilities under this new order. As one panel member stated, "In a sense the buck has passed from the central office to the school office."

Given the findings from this study, we concluded that the elementary principal's expertise in management and administration should continue to expand. It was also concluded that principals in SBM would benefit from staff development programs that provide the opportunity to learn decision-making and management strategies, including time management. Caldwell and Marshall (1982) advise that in a staff development program which focuses on school improvement "it is assumed that if the individually identified needs of professional staff are met within the context of institutional goals, the best possible education can be provided for the students." (p. 33)

Research Question Two

What changes have occurred in the elementary principal's role with respect to leadership after the implementation of SBM? Although consensus was reached on items 17, 20, 16, 15, 19, 21 and 22, only item number 17 achieved better than 75% agreement:

- 17 : The responsibility of the principal has increased to cultivate leadership from the ranks of teachers (85.7%)

The other six items ranged from 66.7% to 71.4%. In their responses concerning both the changes in the role of the principal in management/administration and leadership, the expert panel expressed its frustration in the increased amount of time put forth by site administrators working in SBM.

Experts in this study concurred that the SBM process with shared governance has created a time management problem for administrators. One panel member, a district level administrator, expressed disappointment that "the number of meetings an administrator attends and often orchestrates has increased ten fold in only a few short years." He went on to say, "gathering ideas and suggestions often creates time barriers that slow implementation."

With these findings serving as a basis for support, it can be concluded that the leadership process in SBM has become cumbersome because of the need for information from all of the stake holders. Time to focus on conducting school-based management processes is a critical factor in the success of SBM (Murphy & Beck, 1995). Elementary principals need to develop a comprehensive plan for coordinating groups and meetings. They will also benefit from leadership training.

Research Question Three

What are the primary management/administrative tasks of the elementary principal in SBM? Consensus was reached on 18 items. Items 34, 38, 39, 37, 36, 43, 26, and 32 ranged between 60% and 75% agreement, while the following statements achieved a level of agreement higher than 75%.

- 40 : Promoting the vision and the mission of the school (90.5%)
- 27 : Facilitating the involvement of others in school decision-making (81.0%)
- 29 : Carrying out the ideas developed by the group (81.0%)
- 23 : Building consensus (76.2%)
- 24 : Staying abreast of the work of the whole school while allowing people to assume responsibility for their part (76.2%)
- 30 : Orchestrating meetings (76.2%)
- 31 : Serving as the manager of people at the site level (e.g., Providing for the recruitment selection, development, evaluation and, if necessary, the separation of faculty and staff members who work in the school) (76.2%)
- 33 : Creating organizational structure (e.g., Work teams) for school that involves all faculty members in decision-making (76.2%)
- 35 : Recognizing all "SUCCESSES" (76.2%)

Panelists indicate that the promotion of the vision and the mission was superior to other items related to the elementary principal's primary tasks in management/administration. Bennis (1989) stated that "true leaders work to gain the trust of their constituents, communicate their vision lucidly, and thus involve everyone in the processes of change" (p. 30). Panel members concurred that the elementary principal must function to keep the stake holders focused on the goals set forth in the mission statement. According to theses findings, it may be concluded that strategic planning concepts are vital to SBM. Strategic planning is a tool for rethinking, restructuring, and revitalizing education (Kaufman, Herman, & Waters, 1996).

Research Question Four

What are the primary leadership tasks of the elementary principal in SBM? Statements on which consensus was gained and also ranging above 75% in agreement were:

- 52 : Keeping the staff informed (90.5%)
- 47 : Facilitating the involvement of others in decision making (85.7%)
- 50 : Providing opportunities for professional growth for all staff (85.7%)
- 53 : Communicating with all school constituencies (85.7%)
- 54 : Facilitating the change process (85.7%)
- 48 : Building a school-wide vision (81.0%)
- 49 : Promoting strategic planning for school improvement efforts (81.0%)
- 55 : Organizing meetings (81.0%)
- 58 : Helping the School Leadership Team members to build coalitions for the greater good of all students (81.0%)
- 56 : Overseeing the operation of the school (budgeting, scheduling, hiring, etc.) (76.2%)

Items 45, 57, and 46 (Coaching, Carrying out democratically made decisions, and Building consensus) were the remaining consensus statements. Their level of agreement was below 75%.

According to these findings, highest leadership priority should be given to keeping the staff informed, one of the keys to the success of SBM. "Particularly in a large school, the distribution of information is critical," according to one panelist. Another panelist commented that creating organizational structures whereby all those in the school are involved in decision-making is vital.

In light of the findings of this study, it was concluded that in SBM elementary principals need to work toward becoming master facilitators and communicators. Sound backgrounds in strategic planning and group management are essential.

Research Question Five

How does the implementing of SBM alter the role of the elementary principal in the decision-making process? As noted by the experts in this study, the pervasive idea that principals will negate their power and responsibilities because of SBM is not true. Panelists agreed that principals in SBM retain the authority and responsibility for some decisions. They state, however, that in SBM, the principal has a commitment to the empowerment of teachers in the decision-making process and seeks to give teachers the opportunity to be active in the shared governance undertaking. The findings suggest that the principal, by participating with others in the decision-making process and seeking ways to empower teachers to be responsible for the resolution of instructional issues, has become a leader of leaders.

"Shared decision-making is difficult when the staff continues to be isolated" (Squires & Kranyik, 1996, p. 29). Panelists suggested the principal is responsible for creating organizational structures in the school that involve all faculty members in decision-making. One principal remarked, "I recognize that it is our school, not my school and that synergy produces better solutions to problems than I can figure out by myself."

Inferred from the findings of this study is a need to identify specialized proficiencies essential for leadership support of productive shared decision-making. This suggests that professional development programs for administrators may need to be adapted to accommodate the advancement of new competencies evolving from the implementation of SBM. It also can be concluded from the data that it is a responsibility of the principal to keep constituencies abreast of vital information basic for making informed decisions. Experts in this study noted as a coach, the principal works to create a supportive environment that encourages risk- taking and participation in collaborative decision-making processes. Their perception was that it is becoming increasingly significant for the principal to create a climate in which teacher leadership may evolve. Coordinating the development of a distribution system through which information is provided to decision makers on how to prepare budgets, hire personnel, develop schedules, and plan the curriculum has emerged as an essential role of the principal in SBM, panel members remarked.

Summary

This study was completed to detect the realities and the perceptions of selected administrative functions (leadership, decision making, and management) of the elementary principalship under SBM policy and create a job profile for that position. Given the content, level of agreement, and stability of each of the final 48 items, many conclusions may be made. The examples, as shown below, are drawn from the consensus statements having at least 80% agreement among the experts.

Changes in Administration, Management, and Leadership

After implementation of SBM policy,

- The elementary school principal working in SBM should share the responsibility of attaining the school's goals with all collaborating parties,
- Orchestrate shared decision making,
- Practice time management techniques,
- Obtain knowledge concerning group process and interpersonal skills, and
- Cultivate leadership from the ranks of teachers.

Job Profile

The primary tasks of the elementary principal working under SBM policy with shared governance are to

- Promote the mission of the school,
- Facilitate the involvement of others in school decision making,
- Implement ideas developed by the group,

- Keep the staff informed,
- Encourage the involvement of others in decision making,
- Provide opportunities for professional growth for all staff,
- Communicate with all school constituencies,
- Foster the change process,
- Build a school-wide vision,
- Advance strategic planning for school improvement efforts
- Organize meetings, and
- Help the School Leadership Team to build coalitions for the good of all students.

Recommendations for Further Research

To augment the results of this study and to gain a composite of the elementary school principal's role and primary tasks in implementing SBM, the following recommendations are made for additional research:

1. The results of this study should be expanded to include a comprehensive survey of elementary principals in schools that are implementing SBM at the National level. This study would further define and clarify the roles and tasks of the elementary principal in SBM and validate the findings in this study. The Job Profile of the Elementary School Principal in SBM might be used as part of the survey instrument. Comparisons might be made with the findings of this inquiry and the results of such a study would be beneficial in determining the course of study for principal preparation programs.
2. SBM, as revealed in the literature, requires new skills for the leadership roles and responsibilities of teachers and administrators in elementary schools. However, existing literature does not offer specific data to confirm exactly what professional development practices maximize the effectiveness of SBM. Further studies are needed to assess the effectiveness of professional development programs in elementary schools implementing SBM.
3. Items on which consensus was not reached need further investigation.

Conclusion

This study suggests that within the context of a school working under SBM policy, the elementary principal's role as leader requires specialized skills to support participative management. Considerations need to be made by colleges, universities, and job performance centers to assess their administrative training programs for congruence with changes in the field. Consideration should be given to restructuring traditional educational administration training to include the knowledge and skills indigenous to SBM such as principles of strategic management, facilitating group processes, building consensus, and enabling communications.

References

Aronstein, L. W., & DeBenedictis, K. L. (1991). An interactive workshop: Encouraging school-based management. NASSP Bulletin, 75(537), 67-72.

Bennis, W. (1989). *Why leaders can't lead.* San Francisco, CA: Jossey-Bass Publishers.

Black, S. (1996). Share the power. *The Executive Educator*, 18(2), 24-26.

Blase, J. (1987). Dimensions of effective school leadership: The teacher's perspective, *American Educational Research Journal*, 24(4), 589-610.

Borg, W. R., & Gall, M. D. (1983). *Educational research: An introduction* (4th ed.). New York: Longman.

Caldwell, S. D., & Marshall, J. C. (1982). Staff development--four approaches described, assessed for practitioner, theoretician. NASSP Bulletin, 66(451), 25-35.

Caldwell, S. D., & Wood, F. H. (1992). Breaking ground in restructuring. *Educational Leadership*, 50(1), 41-44.

Carlin, P. M. (1992). The principal's role in urban school reform. *Education and Urban Society*, 25(1), 45-56.

Conway, J. A., & Calzi, F. (1996). The dark side of shared decision making. *Educational Leadership*, 53(4), 45-49.

Cyphert, F., & Gant, W. (1971). The Delphi technique: A case study. *Phi Delta Kappan*, L11(3), 272-273.

Daniels, C. T. (1990). A principal's view: Giving up my traditional ship. *The School Administrator*, 47(8), 20-24.

David, J. L. (1996). The who, what, and why of site-based management. *Educational Leadership*, 53(4), 4-9.

Doud, J. L. (1989). The K-8 principal in 1988. *Principal*, 68(3), 6- 12.

Drury, W. R. (1993). The principal's role in site-based management. *Principal*, 73(1), 16-19.

Fiske, E. B. (1995). Systematic school reform: Implications for architecture. In A. Meek (Ed.) *Designing Places for Learning*. (pp. 1-10). Alexandria, VA: ASCD.

Gleason, S. C., Donohue, N., & Leader, G. C. (1996). Boston revisits school-based management. *Educational Leadership*, 53(4), 24-27.

Goldman, P. (1991, April). Administrative facilitation and site- based school reform projects. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Documentation Reproduction Service No. ED 332 334)

Goodman, J. (1994). External change agents and grassroots school reform: Reflections from the field. *Journal of Curriculum and Supervision*, 9(2), 113-135.

Guskey, T. R., & Peterson, K. D. (1996). The road to classroom change. *Educational Leadership*, 53(4), 10-14.

Hallinger, P., Bickman, L., & K. Davis (1990). What makes a difference? School context, principal leadership, and student achievement. (Occasional Paper No. 3). Cambridge, MA: The National Center for Educational Leadership.

Harrison, C. R., Killion, J. P., & Mitchell, J. E. (1989). Site-based management: The realities of implementation. *Educational Leadership*, 46(8), 55-58.

Helmer, O. (1967). *Analysis of the Future: The Delphi Method*. Santa Monica, CA: Rand Corporation.

Kaufman, R., Herman. J., & Waters, K. (1996). *Educational Planning*. Lancaster, PA: Technomic.

Krug, S. E. (1993). Leadership craft and the crafting of school leaders. *Phi Delta Kappan*, 75(3), 240-245.

Lindelow, J. (1981). School-based management. School Management Digest, (Series 1, No. 23). (ERIC Documentation Reproduction Service No. ED 208 452)

Linstone, H. A., & Turoff, M. (Eds.). (1975). *The Delphi method: Techniques and applications*. Reading, MA: Addison-Wesley.

Murphy, J., & Beck, L. G. (1995). School-based management--taking stock. *Kappa Delta Pi Record*, 32(1), 6-10.

Putnam, J. W., Spiegel, A. N., & Bruininks, R. H. (1995). Future directions in education and inclusion of students with disabilities: A Delphi investigation. *Exceptional Children*, 61(6), 553-576.

Sackman, H. (1974). *Delphi assessment: Expert opinion, forecasting, and group process*. Santa Monica, CA: Rand Corporation.

Spilman, C. E. (1996). Transforming an urban school. *Educational Leadership*, 53(4), 34-39.

Squires, D. A., & Kranyik, R. D. (1996). The Comer program: Changing school culture. *Educational Leadership,* 53(4), 29-32.

Stephens, E. R. (1987). Improving the effectiveness of school-based administration in Maryland. Baltimore, MD: Maryland Commission on School-Based Administration. (ERIC Document Reproduction Service No. ED 283 237)

Tanner, C. K., & Williams, E. J. (1981). *Educational planning and decision making.* Lexington, MA: Lexington Books.

Thurston, P., Clift, R., & Schacht, M. (1993). Preparing leaders for change oriented schools. *Phi Delta Kappan,* 75(3), 259-266.

Uhl, N. P. (Ed.). (1983). *Using research for strategic planning.* San Francisco: Jossey-Bass.

Vann, A. S. (1996). An alternative assessment for master teachers. *Principal,* 75(3), 29-30.

Wohlstetter, P., & Briggs, K. L. (1994). The principal's role in school-based management. *Principal,* 74(2), 14-17.

Wohlstetter, P., & Odden, A. (1992). Rethinking school-based management policy and research. *Educational Administration Quarterly,* 28(4), 529-549.

Appendix

National Expert Panel Summary Information

Names were not included in order to preserve anonymity. Locations, occupational positions, and panel nomination sources of the selected panel members are stated, and reference is also made to the qualifications of the panel members as experts in SBM.

The composition of the panel originally consisted of 24 panel members, 13 males and 11 females. Eleven males and 10 females comprised the panel at the end of three cycles.

One of the objectives of the panel selection process was to select SBM experts that represented various regions across the United States. Of the original 24 panel members, two principals and two specialists were from the Pacific Coast States of California and Washington. One principal and one specialist were located in the Southwest Region in the state of Texas. The Heartland, comprised of Missouri and Nebraska, was represented by two principals and one specialist. Four specialists and four principals resided in the Southeast Region States of Kentucky, Florida and Georgia. The Mid-Atlantic area was represented by three principals and one specialist from Maryland, New Jersey, New York, and Pennsylvania. Two specialists were located in the Great Lakes area of Indiana and Ohio.

Various educational occupations were represented by expert panel members: (a) principals, (b) assistant superintendents, (c) director of a center for educational governance, (d) book authors, (e) lecturer and author on school reform. (f) director of a school improvement organization, (g) a Governor's Leadership Institute consultant, (h) director of a center for leadership development, (i) consultant for a performance improvement corporation, (j) director of school principals, (k) creator of a principal's training center, (l) retired chair of a department of educational administration, (m) staff members of leadership training institutes, and (n) area superintendents.

Names of the selected panel members who were principals were obtained from the following sources: (a) Two principals were nominated by a Dean of the College of Education at a large university. The school is involved in the development of educational governance. (b) Two panel members were honored as nationally distinguished principals. (c) Five principals or their schools had been published, cited, or recognized in a nationally distributed journal. (d) five principals were recommended by the Superintendent's office of school districts involved and/or cited in SBM literature. (e) Two principals were nominated by university professors who had published articles on SBM in nationally distributed journals. (f) One principal was nominated by the Director of a university program involved in school reform. (g) Two principals were National Association of Elementary School Principals (NAESP) Distinguished Principals.

Justification for the specialists chosen to serve on the panel was based on the following criteria: (a) Five had published in nationally distributed journals. (b) The school districts of three specialists had been cited in the SBM literature. These specialists were administrators in these districts and were involved in the district's implementation of SBM. (c) Four specialists were Directors or staff members of leadership development centers supportive of SBM with shared governance and shared decision-making. (d) Two specialists were administrators in school improvement organizations. (e) Two specialists have written books relative to school reform, school improvement, and educational leadership. (f) Two specialists had presented research papers at a meeting of the American Educational Research Association. (g) Three specialists were involved with their own leadership improvement corporations.

Principals on the panel had published in the following nationally distributed journals: The Executive Educator, Principal, The School Administrator, and Educational Leadership. Specialists on the panel had published in the following periodicals: Educational Administration Quarterly, NASSP Bulletin, and Principal.

About the Authors

C. Kenneth Tanner, Professor

Department of Educational Leadership

310 River's Crossing
The University of Georgia
Athens, GA 30602

Email: ktanner@coe.uga.edu

Phone (706) 542-4067
Fax (706) 542-5873

Dr. Tanner serves as Professor of Educational Leadership, the University of Georgia. His primary research interests are in the fields of educational policy analysis and school design and planning. Dr. Tanner has published three books on planning and written over 100 articles, papers, and chapters, which deal with policy and planning. His recent planning activities may be found at the SDPL's Web site: http://www.coe.uga.edu/sdpl/sdpl.html

Cheryl D. Stone, Principal

M. G. Barksdale Elementary School
Rockdale County Schools
Conyers, GA 30208-4199

Email: drcherry@aol.com

Phone (770) 483-9514
Fax (770) 483-0665

Dr. Stone is the Principal of M. G. Barksdale Elementary School in Conyers, Georgia. She is a proponent of site-based management and has facilitated the resolution of school policies and decisions through shared governance processes modeled as an "umbrella style" organizational decision-making structure. Barksdale was awarded the 1998 Connecting Teachers with Technology Grant from USWest/Media One and has been honored as a 1995 Georgia School of Excellence. Barksdale is a member of the League of Professional Schools.

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

Greg Camilli
Rutgers University

John Covaleskie
Northern Michigan University

Andrew Coulson
a_coulson@msn.com

Alan Davis
University of Colorado, Denver

Sherman Dorn
University of South Florida

Mark E. Fetler
California Commission on Teacher Credentialing

Richard Garlikov
hmwkhelp@scott.net

Thomas F. Green
Syracuse University

Alison I. Griffith
York University

Arlen Gullickson
Western Michigan University

Ernest R. House
University of Colorado

Aimee Howley
Marshall University

Craig B. Howley
Appalachia Educational Laboratory

William Hunter
University of Calgary

Richard M. Jaeger
University of North Carolina--Greensboro

Daniel Kallós
Umeå University

Benjamin Levin
University of Manitoba

Thomas Mauhs-Pugh
Rocky Mountain College

Dewayne Matthews
Western Interstate Commission for Higher Education

William McInerney
Purdue University

Mary P. McKeown
Arizona Board of Regents

Les McLean
University of Toronto

Susan Bobbitt Nolen
University of Washington

Anne L. Pemberton
apembert@pen.k12.va.us

Hugh G. Petrie
SUNY Buffalo

Richard C. Richardson
Arizona State University

Anthony G. Rud Jr.
Purdue University

Dennis Sayers
University of California at Davis

Jay D. Scribner
University of Texas at Austin

Michael Scriven
scriven@aol.com

Robert E. Stake
University of Illinois--UC

Robert Stonehill
U.S. Department of Education

Robert T. Stout
Arizona State University

---

archives | abstracts | editors | board | submit | comment | subscribe | search

EPAA
- enter the archives
- browse the abstracts
- the editors
- the edit board
- submit article
- submit commentary
- search
- subscribe

volume: ● 1 ● 2 ● 3 ● 4 ● 5 ● 6 ● 7

This article has been retrieved 737 times since March 30, 1998.

Education Policy Analysis Archives

Volume 6 Number 7 **March 30, 1998** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal.
Editor: Gene V Glass Glass@ASU.EDU.
College of Education Arizona State University,Tempe AZ 85287-2411

Copyright 1998, the EDUCATION POLICY ANALYSIS ARCHIVES.Permission is hereby granted to copy any article provided that EDUCATION POLICY ANALYSIS ARCHIVES is credited and copies are not sold.

Educational Research in Latin America: Review and Perspectives

Abdeljalil Akkari
University of Maryland Baltimore County

Soledad Perez
Geneva University

Abstract

The present paper consists of four primary sections. First, we describe the historical context of educational research in Latin America. In the second section, we focus on various theoretical frameworks that are applied to educational research in the region. We identify the main institutions involved in this research in the third section. Finally, in conclusion we offer suggestions that we consider to be of greatest priority for the future of educational research in Latin America.

Introduction: Historical context

During the 1930s, many institutions in Latin American countries began to conduct research in the area of education. Shared historical and political ties as well as similarities among educational systems within the region helped facilitate comparative and international studies in Latin America . Since 1958, the Organization of

Iberoamerican States (OIS) started to develop numerous studies in different countries, with many publications stemming from this work. However, most research has primarily been descriptive rather than empirical or applied. We found that when strong coordination exists between the local governments and international agencies, the results are more relevant and useful. One such effort is the report, "The demographic, economic, social, and educational situation in Latin America," (United Nations Educational, Scientific, and Cultural Organization or UNESCO, 1962) conducted by the Organization of American States (OAS), the Food and Agriculture Organization of the United Nations (FAO), the International Labor Organization (ILO), and the Economic Commission for Latin America and the Carribean (ECLAC).

During the 1970s, the areas of educational research and economic development began to join forces. An example of this union was a program on functional literacy, conducted by UNESCO. Researchers started to evaluate their own work and to examine the effects that their research had on educational development. They found that universal schooling for all children did not exist and that in many ways societies continued to suffer from social inequality, as was the case before governments began to invest in basic education. However, due to the dictatorial atmosphere in Latin America during this period, researchers were limited in conducting innovative research, especially in the realm of education.

Multiple diversification of research began to take place during the period of the '80s and '90s. A wave of democratization across the region acted as a catalyst for this diversification. First, regarding topics and methodology, a stronger emphasis was placed on the disciplines of anthropology, sociology, and psychology. Many groups, such as indigenous populations who were previously ignored by mainstream research, became the subjects of investigation. In addition, there were many new institutions involved in educational research, specifically non- governmental and international agencies.

Presently, the scope of educational research in Latin America is great, including didactic methods of teaching, non-formal education, and adult literacy. However, as suggested by Tedesco (1995), progressive theoretical frameworks and previously implemented educational programs are not alone sufficient in delineating an educational orientation capable of obtaining the goals of democracy and the overall equal distribution of knowledge.

Theoretical Frameworks

While there exist many different theoretical frameworks, we will focus primarily on those that have played an influential role among Latin American educators and researchers. During the period from 1950-1965, the main theoretical framework was based on studies conducted by ECLAC. This agency encouraged Latin America to transform from a traditional society dominated by large land proprietors (latifundios) to a modern one based on productivity and growing industrialization.

In viewing education as a resource for economic development, the agency attempted to apply human resources theory (Shultz, 1981) to countries in Latin America. According to this theory, education is considered primarily as an investment in human capital, with substantial long- term benefits both for the individual being educated and for the community as a whole. During this period the main debate centered around accessibility to schooling and duration of attendance for those children from impoverished and rural areas. In addition, public funding and teacher training were topics of interest during this time (Garcia- Huidobro, 1990).

It is important to recognize that during this period a strong ideological debate was taking place in Latin America, with the central hope being that the Cuban revolution would bring radical social transformation. In reaction to ECLAC s theoretical model, described above, was a more progressive and political approach, one organized around dependency theory (Cardozo & Faletto, 1969). Dependency theory focuses on a macro level of socioeconomic change rather than on an individual level. It is based on the idea that third world societies depend economically on industrialized countries. There is also an element of internal domination that exists between different socioeconomic groups within each country. Dominant groups attempt to perpetuate the situation of inequality by controlling systems of production and education. However, research linked to this theory does not systematically apply to issues of education.

A third theory, closely linked with the pedagogical work of Freire (1972), emerged during the late '60s and early '70s. While partially influenced by both Liberation Theology and Illich's work (1973), Freire suggested that the main goal of education was "liberation." According to this objective, those individuals involved in education must shift from the status of object (being taught) to the status of actor/subject (learning). According to Freire (1972, 1976), effective education includes theoretical as well as practical knowledge that relates to the local context rather than decontextualized curricula. Empowerment is both the means and the outcome of Freire's pedagogy, which some have come to call "liberatory education."

Another central concept in Freire's work is that traditional Brazilian education is based primarily on "cultural hegemony" (1972). Similarly, Carnoy (1974) proposed a historical critique of education by defining it as "cultural imperialism." In examining the history of education in Latin America, we concur with Freire in that mainstream education in the region works to maintain social, political, and economic domination of subordinate groups. We also agree with Freire's belief that the role of schooling in Latin America contributed to maintaining the poor at the bottom of the social structure. He strongly believed that society shapes school rather than school shaping society. In addition, Freire's involvement was mainly with adult literacy classes. Because the scope of Freire's theory encompasses primarily adult education issues, its pedagogical application is limited in understanding the dynamics of formal schooling in Latin America.

The debate around this theory led scholars to adopt participatory research as a part of their work on education during the 1970s. For instance, Meister (1968) explored how rural populations utilized education to obtain greater access to economic development. In Ecuador, with the help of the Catholic church, local communities created novel radio programs to promote adult literacy. These people became active participants in their education through their involvement in both the conceptualization and the realization of these radio programs.

During the early 1980s, ECLAC proposed a new slogan, "Productive transformation with equity," one that paralleled the increasing democratization of local regimes. There was a shift in focus toward pragmatic goals rather than on global transformation, as previously suggested by ECLAC during the '50s. A key issue involved in this new pragmatism was the priority to narrow the gap between education and the work force. A network was created in the region to identify the relation between education and work (Red Latin American Education and Work). Within this network, there was a great deal of discussion regarding the meaning of both education and work in relation to different socioeconomic groups (De Ibarrola, 1990; Filmus, 1995). In summary, the two main theoretical influences come from ECLAC as well as from remnants of Freire's work. While ECLAC focuses on the quantitative side of improving basic education, Freire s theory relates well to adult literacy but remains limited when applied to formal schooling.

Sources of Institutional Research

This section does not provide an exhaustive description of all institutions involved in educational research throughout Latin America, but instead focuses on exemplary cases. In most of the countries, we will distinguish between four sources of institutional research: a) state agencies, b) universities, c) non-government organizations (local, international, religious, etc.), and d) foreign aid agencies.

State agencies

Most countries in Latin America have specific agencies that are responsible for educational research. These agencies are typically related to the Ministry of Education. An example is the National Institute of Educational Studies and Research (INEP), previously named the National Institute of Pedagogical Studies, created in Brazil in 1938. Since its inception, the institute has been responsible for the continual publication of the Brazilian Journal of Educational Studies, one of the main sources of information and analysis regarding Brazilian education. At a research level, INEP supported an average of 24 research projects a year from 1972 to 1982. This is an insufficient number in comparison to the need of the population of 160 million inhabitants.

In the region as a whole, state agencies played a dynamic role in educational research during the '60s and '70s. However, this funding and support gradually decreased over time. As a first measure the

National Institute of Educational Research (INIDE) in Peru was initially the target of budget reduction and was subsequently closed. While in 1977 there were 90 research projects under way, by 1985, there were only 50 projects remaining. Presently, the National Agency of Educational Research and Teacher Training (DINIC) coordinates educational research mainly on psycho- pedagogy and the sociology of education. This same situation occurred in Chile, with support for the Center of Improvement, Experimentation, and Education Research (CDEIP) being severely cut in Santiago. Similarly, in 1968 the Colombian Institute of Pedagogy (ICOLPE) was closed and replaced by the Colombian Fund of Scientific Investigations and Special Projects (COLCIENCIAS).

Those researchers involved in Latin America agree that there is no need to create more institutions but rather to increase coordination among the already existing institutions (Pizzarro, 1990). There are various new centers that provide resources as well as help to facilitate research through offering access to technological information such as databases and on-line publications. An example of such a center is the Caribbean Research Information Service (CERIS), located in Trinidad and Tobago. This center focuses on four primary goals (Velloso, 1996):

1. gathering information related to the structure of educational systems
2. collecting information related to current research on education
3. constructing an annotated bibliography on completed research
4. identifying potential sources of information

Universities

In many cases, Education Departments are responsible for the majority of research activities. Currently in Brazil, universities conduct approximately 80% of the research, while the remainder is carried out by non- governmental organizations and municipal agencies. However, previously there was a paucity of experienced researchers in Brazil; it was not until the late '70s that graduate programs in education were developed.

In 1981, 22 different universities offered a total of 27 graduate programs in education (549 masters theses and 10 doctoral dissertations). Velloso (1996) suggests that we have had a quantitative gap during the last decade, with 4,000 masters theses and 400 doctoral dissertations completed in the following universities: the Federal Universities of Minas Gerais, Rio Grande do Sul, Rio de Janeiro, and Fluminence; the State Universities of Sao Paulo and Campinas; the Catholic University of Sao Paulo. In Chile, the Education Departments in the Catholic University of Chile and the University of Chile conduct research primarily on educational policy and planning. In the Caribbean region, research in education was initiated in 1954 with the birth of the Educational Research Center, affiliated with the West Indies University.

Due to the increasing number of students enrolled in education programs throughout Latin America, professors have had to focus more on teaching rather than research. In addition, professors are able to obtain tenure once they have completed their masters degree, and without mandatorily proceeding with any doctoral research. During the last decade, there have been many initiatives aimed at increasing both the quantity and the quality of educational research and to promote communication between scholars. For the past fifteen years, the National Brazilian Association for Graduate Studies in Education (ANPE) has held annual meetings as well as regular workshops on specific topics where researchers come together to discuss and share their current work on education. In addition, there are two federal organizations in Brazil, the National Council of Scientific and Technological Development (CNPq) and the Commission of Advancement and Training of University Personnel (CAPES), that send students abroad to study in doctoral programs and evaluate local graduate programs. In Chile, the Ministry of Education initiated the Program for the Improvement of Quality and Equity of Education (MECE), a project that combines the efforts of policymakers and researchers and encourages more applied projects. With the help of the World Bank, Paraguay also adopted this initiative to improve the quality of secondary schools. Finally, on a more regional level, there is an academic organization, the Latin American University of Social Sciences (FLACSO), that supports cooperative research among universities throughout Latin America.

Non-governmental organizations (NGOS): Local, international, religious

During the last decade, nongovernmental organizations (NGOS) have played a greater role in educational research in Latin America. In Chile, for instance, NGOS are the main resource for studies in the area. The Research Center on Educational Development (CIDE), founded by the Catholic Church in 1964, presently includes approximately twenty researchers with doctoral degrees who are working on projects that are funded by both local and foreign agencies. This center is responsible for coordinating the Latin American Educational Information and Documentation Network (REDUC), an organization that collects and disseminates periodical information on educational research in seventeen different countries (Analytical Abstracts on Education - RAE). There is a similar NGO in Mexico, the Center of Educational Studies (CEE), that offers the most complete database on educational research in Mexico. In Brazil, the Carlos Chagas Foundation in Sao Paulo is involved in both traditional as well as more innovative research projects, such as a project on ethnicity and education. In addition, this foundation publishes an international journal, Cadernos de Pesquisa.

Multinational corporations have traditionally been involved in funding educational projects, but more recently they have begun to participate in new projects that also include a research dimension. For example, the Swiss Foundation for Sustainable Development (NOVARTIS) provides funding for and monitoring of community centers for street children. Another example can be found in a distance learning program, "Telecurso 2000," sponsored by the Roberto Marinho Foundation, a project based around three primary goals: contextual teaching (ensino em contexto), development of fundamental competency, and citizenship empowerment. Latin American NGOS receive the majority of their funding from foreign sources. CIDE received 60% of its funding from sources outside of Chile (Velloso, 1996). Similarly, the Carlos Chagas Foundation in Chile received many contributions from the Ford Foundation. The International Center for Research on Development (ICRD), a Canadian organization, offered substantial funding to local Latin American NGOS during the last twenty years. One research project that was recently funded investigates survival strategies for marginalized groups in the workplace in Uruguay (Lemez, 1997).

Foreign aid agencies

Regional and international foreign aid agencies have participated in educational projects since the '50s. One project, Development and Education in Latin American and the Caribbean (UNDP, 1981), supported by three different international agencies, ECLAC, UNESCO, and the United Nations Development Program (UNDP), offered a comparative perspective of education among the countries in the region. More recently, the World Bank has come to play a predominant role in financing educational projects in the region (Coraggio, 1995). Between 1990 and 1994, the World Bank was responsible for contributing $1.1 billion annually to educational projects in Latin America (MacMeekin, 1996). Proceeding the World Bank, the three principal funding agencies are the Inter-American Development Bank (IDB), the Japanese Agency of International Aid, and the USA International Agency (USAID).

The Nordeste Project in Brazil is a typical example of the World Bank's involvement in education. It is the largest investment of the World Bank in Brazil. During the next four years, $736.5 million will be allocated to improve the quality of primary school education (ensino fundamental de 1 a 4 seria). Due to the fact that 30% of children between seven and fourteen years are not attending school and that only 76% of adults have less than four years of schooling explain the need for such an investment in education (MEC, 1996). However, in examining the general orientation of this project, we do not believe that it will improve the local educational situation. First, the project does not have a sufficient research team that would be able to properly evaluate relevant problems in local schools. The philosophy underlying the project lacks a global understanding of the relationships between education and society, and instead focuses on fragmented and quantitative goals within the formal schooling system. The overriding theme of the World Bank's Nordeste Project revolves around material and human management. While one third of the funds have already been allocated toward the purchase of 47 million textbooks (Sanjuro, 1996), these materials are primarily designed and manufactured in industrialized southern Brazil, a context very different from the local rural one. While the project s main focus is education, we believe it also needs to consider and incorporate the fact that land is unequally distributed within the surrounding region. However, as suggested by Coraggio (1995), the redistribution of productive resources among different socioeconomic groups is not a priority of the World Bank.

By examining different sources of institutional research, we are able to identify some general tendencies that are common across Latin America. First, while state agencies previously had a substantial role in educational research, they are presently downsizing their research capacity. Second, universities have maintained their position, especially in relation to countries in which there is a strong academic tradition. Finally, with contributions from private and public foreign agencies, NGOS have increased their research capacity.

Future Perspectives

Highlighted in both electoral campaigns and official documents, education seems to be the main priority in many Latin American countries. However, in viewing the actual educational situation in the region as a whole, we can see that there still remains a great deal more that needs to be accomplished. Central America is particularly affected, with 1.5 million children still outside of the school system. Indeed, the previous objective of generalization of primary education has yet to be reached, especially for indigenous children, poor children, and those living in rural areas. As pointed out by Puryear (1996, p.3), "Latin America's primary and secondary schools are sharply segmented by economic status, with the poor consigned to the public system while the rich and most of the middle-class attend private schools." As previously stated, there is a need for a new approach to education in Latin America that supersedes both the human resources theory supported by ECLAC as well as Freire's pedagogy of adult literacy.

The schooling system is structurally divided into two separate networks: private and public. This division parallels Baudelot's and Establet's (1971; 1975) conceptualization of the French school system. The separation found in their work is between primary education-professional training and secondary education-university. Children from working class families are confined to the first network while children from an upper-class background are likely to reach levels of secondary and university education. Another parallel can be found in Serpell s analysis of schooling in Africa. Serpell argues that, "...the narrowing staircase model of schooling, which informs the prototype of Institutionalised Public Basic Education (IPBS) combines a metaphor of the individual s developmental progress as climbing a staircase with a conception of the social function of schooling as the recruitment of an elite by gradually extracting them from humble origins into a privileged upper class" (Serpell, in press). In this way, schooling in many third world countries functions in an "extractive manner," working against "the principle of local accountability in both the economic and the cultural sphere."

We see that in as early as the primary school years, there already exists a fundamental qualitative difference between public and private education in Latin America. We can easily observe that classrooms are overcrowded and that teachers have minimal training in many public schools. Numerous children drop out of the system prematurely and some of them later go on to some type of nonformal education.

We also observe a strong residential segregation in Latin American countries. Regions are made up of distinct socioeconomic groups that are very different from each other and children within the public school sector have very little opportunity to interact outside of school with children who attend private schools. It is as if there exist two parallel processes of socialization, and this poses a challenge for those citizens who live together and are working toward a common future. Hence, one main goal of educational research in Latin America is to first investigate the existing segregation and later reconstruct a new model of public education. Accomplishments such as these will be relevant to educational issues in other third world countries because of the increasing deterioration of the quality of public education and the tendency for children from upper-class families to obtain private rather than public education.

References

Baudelot, & Establet. (1971). *L'cole capitaliste en France*. Paris: d. Maspero.

Baudelot, & Establet. (1975). *L'cole primaire divise*. Paris: d. Maspero.

Cardozo, F. H., & Faletto. (1969). *Dependencia y desarrollo en America Latina*. Mexico: Siglo XXI.

Carnoy, M. (1977). *La educacion como imperialismo cultural*. Mexico: Siglo XXI.

Coraggio, J. L. (1995). Educacion y modelo de desarrollo. In CEAAL (Ed.), *Construccion delas politicas educativas de Americas Latina*, (pp. 83-131). Lima: Tarea.

De Ibarrola, M. (1990). Projecto socioeducativo, institution escolar y mercado de trabajo: el caso del Tecnico Medio Agropecuario. Unpublished Doctorado, UNM, Mexico.

Freire, P. (1972). *The pedagogy of the oppressed*. Harmondsworth: Penguin.

Freire, P. (1976). L'alphabetisation et le "reve possible". *Perspectives*, vol VI, pp. 70-73.

Garcia-Huidobro, J., Tellet, F. & Ochoa, J. (1990). Tendencia de las investigacion educacional en America Latina. Santiago de Chile:

CIDE..

Illich, I. (1973). *Deschooling society*. Harmondsworth: Penguin.

Lemez. (1989). La educacion y las strategias sociales de sobrevivencia en el mercado de trabajo. Montevido: Centro d'investigacion y Experimentacion Pedagogica (CIEP).

McMeekin, R. (1996). Coordinacion de la asistencia externa para la educacion en America Latina y el Caraibe. *Boletin del Projecto Principal de Educacion*, abril(39), 20- 54.

MEC. (1996). Projecto Nordeste. Brasilia: Ministerio da Educacacao e do Desporto.

Meister, A. (1968). *Participation, animation, dveloppement*. Paris: Ed. Anthropos.

Pizarro, J. (1990). Investigacion de la Educacion en Algunos Paises de America latina . Ottawa: Centre international de recherche pour le dveloppement (CIRD).

UNDP (1981). Dveloppement et ducation en Amrique latine et les Carabes. Paris: Unesco

Puryear, J. M. (1996). Education in Latin America: Problem and Chalenges (Working Group on Educational Reform). New York: Concil on Foreign Relations.

Sanjuro, R. (1996). Distribuicao do livro didactico no Nordeste brasileiro (Projecto Nordeste). Brasilia: MEC.

Schultz, T. W. (1981). *Investing in people. The economics of population quality*. Berkley: University of California Press.

Serpell, R. (in press). Local accountability to rural communities: A challege for educational planing in Africa. In F. Leach and A. Little (Eds.), *Schools, Culture and Economics in the Developing World: Tension and Conflict*. New York: Garland.

Tedesco, J. C. (1995). *El nuevo pacto educativo. Educacion, competitividad y ciudadania en a sociedad moderna*. Madrid: Grupo Anaya.

UNESCO. (1962). Situacion demografica, economica, social y educativa en America Latina. Paris: UNESCO/CEDES.

Velloso, J. (1996). SERI and capacity building in educational research in Latin America and the Caribbean. In SERI (Ed.), *Educational research in the South: An Initial Review* . Paris: International Institute for Educational Planning.

About the Authors

Abdeljalil Akkari

University of Maryland Baltimore County
Email: akkari@umbc2.umbc.edu

Dr. Abdeljlil teaches courses in multicultural education and sociology of education at Fribourg University (Switzerland). He developed research in Tunisia and Brasil on the relationships between education and development. He is currently as a visiting professor in the Psychology Department of the University of Maryland Baltimore county.

Soledad Perez

Geneva University
Switzerland

Email: perezs@fapse.unige.ch

Dr. Soledad Perez teaches comparative education at the University of Geneva (Switzerland). She is a specialist of Latin American educational systems. She has done research on rural communities in Equator. She collaborated with international organizations on several projects related to education in Latin America, Europe and Africa.

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

Greg Camilli
Rutgers University

John Covaleskie
Northern Michigan University

Andrew Coulson
a_coulson@msn.com

Alan Davis
University of Colorado, Denver

Sherman Dorn
University of South Florida

Mark E. Fetler
California Commission on Teacher Credentialing

Richard Garlikov
hmwkhelp@scott.net

Thomas F. Green
Syracuse University

Alison I. Griffith
York University

Arlen Gullickson
Western Michigan University

Ernest R. House
University of Colorado

Aimee Howley
Marshall University

Craig B. Howley
Appalachia Educational Laboratory

William Hunter
University of Calgary

Richard M. Jaeger
University of North Carolina--Greensboro

Daniel Kallós
Umeå University

Benjamin Levin
University of Manitoba

Thomas Mauhs-Pugh
Rocky Mountain College

Dewayne Matthews
Western Interstate Commission for Higher Education

William McInerney
Purdue University

Mary P. McKeown
Arizona Board of Regents

Les McLean
University of Toronto

Susan Bobbitt Nolen
University of Washington

Anne L. Pemberton
apembert@pen.k12.va.us

Hugh G. Petrie
SUNY Buffalo

Richard C. Richardson
Arizona State University

Anthony G. Rud Jr.
Purdue University

Dennis Sayers
University of California at Davis

Jay D. Scribner
University of Texas at Austin

Michael Scriven
scriven@aol.com

Robert E. Stake
University of Illinois--UC

Robert Stonehill
U.S. Department of Education

Robert T. Stout
Arizona State University

archives | abstracts | editors | board | submit | comment | subscribe | search

EPAA
• enter the archives • browse the abstracts • the editors • the edit board
• submit article • submit commentary • search • subscribe
volume: • 1 • 2 • 3 • 4 • 5 • 6 • 7

This article has been retrieved 1152 times since April 19, 1998.

Education Policy Analysis Archives

Volume 6 Number 8 **April 19, 1998** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal.
Editor: Gene V Glass Glass@ASU.EDU.
College of Education Arizona State University,Tempe AZ 85287-2411 Copyright 1998, the EDUCATION POLICY ANALYSIS ARCHIVES.Permission is hereby granted to copy any article provided that EDUCATION POLICY ANALYSIS ARCHIVES is credited and copies are not sold.

"The Art of Punishing": The Research Assessment Exercise and the Ritualisation of Power in Higher Education

Lee-Anne Broadhead
University of Bradford (U.K.)

Sean Howard
Acronym Institute (U.K.)

Abstract

In this article it is argued that the recent Research Assessment Exercise (RAE)--undertaken by the United Kingdom's Higher Education Funding Councils (HEFC)--is part of a much larger process of assessment in education generally. By taking the RAE as its focus, this article uses a Foucaultian analysis to amplify the nature and practice of disciplinary power in the setting of Higher Education. Foucault's notion of an "integrated system" of control and production, with its routine operation of surveillance and assessment--and its dependence on coercion and consent--is directly applied to the RAE. The impact on research and teaching is discussed. The critical response of academics to the exercise has failed to challenge the process in any fundamental way. it is argued here that this failure is a reflection of the degree to which disciplinary logic is embedded in the academic system.

Introduction

The demands made to "publish or perish" have long played a central role in the academic's career advancement and critiques of this phenomenon are not new. the articulation of a "Research Assessment Exercise" (RAE) within British Higher Education takes this demand to an extreme limit and uses the funding of university Departments as its ultimate weapon. Witnessing the operation of the Exercise has provided a salutary lesson in the effects of product driven research and raises a number of questions about the nature and purpose of academia. The largely unquestioned acceptance of the imposition of such a funding mechanism is a dangerous practice. Positive (if any) and negative effects should be considered. Of greater import academics should reflect on the larger process of which the RAE is a part.

This article provides a critical examination of the use of assessment practices in higher education as exemplified by the RAE. Drawing on Foucault's work on the nature and practice of disciplinary power, the bulk of our consideration will be of the RAE in its internal operation as a ritualisation of such power: a consideration, in Foucault's terms, of the "micro- physics" of power. Naturally, this "micro" level cannot be properly understood without reference to the broader context of British education policy, itself located in a socio-economic setting increasingly characterised by management-centred disciplinarian approaches.

Foucault considers disciplinary power in the context of an "integrated system" of control and production; a system in which, due to the intense, routine operation of surveillance and assessment, both coercion and consent feature prominently. This article investigates the current intensification of the operation of the integrated system in the surveillance and assessment of British academics.

Foucault, Power and the Integrated Disciplinary System

The writing of French philosopher Michel Foucault constitutes one of the most thorough investigations into the evolution and operation of disciplinary systems in the West and the mechanics of power at work within them. Foucault's concern is with the ongoing legacy of "a historical transformation: the gradual extension of the mechanisms of discipline throughout the seventeenth and eighteenth centuries, their spread throughout the whole social body, the formation of what might be called in general the disciplinary society." (Foucault, 1977: 209) The following analysis of the RAE draws most heavily on his 1975 work Discipline and Punish: The Birth of the Prison. As Foucault makes clear, though his study is of prisons, it cannot be a study solely of them, so integrated are they with other forms and elaborations of disciplinary power:

> "'Discipline' may be identified neither with an institution nor with an apparatus; it is a type of power, a modality for its exercise, comprising a whole set of

> instruments, techniques, procedures, levels of application, targets; it is 'physics' or an 'anatomy' of power, a technology. And it may be taken over either by 'specialized' institutions...or by institutions that use it as an essential instrument for a particular end...or by preexisting authorities that find in it a means of reinforcing or reorganizing their internal mechanisms of power..." (Foucault, 1977: 215)

Foucault sees the system of disciplinary power as productive and integrated. He argues that such power cannot rely exclusively or pre-dominantly on punitive measures, essential though these are. For power to be self-sustaining, it must produce and reproduce definitions of reality which the objects of this power come to see as normal. Thus, the moulding and integration of 'the individual' is a central part of the production of power. "Discipline," Foucault argues, "'makes' individuals; it is the specific technique of a power that regards individuals both as objects and as instruments of its exercise." (Foucault, 1977: 170)

In the context of the prison, this "exercise" is designed to be continuous and relentless. Surveillance is the key technique, both of observation and normalisation of behaviour: it integrates the individual within the prison system, "producing" the prisoner, whose ideal variant is highly co-operative and responsive to the authorities. This co-operation is essentially a combination of habitualised, normalised fear of punishment and hope of reward.

In the context of education--identified by Foucault as one of the key sites of the habitualising, normalising exercise of disciplinary power--the primary techniques remain the deployment of surveillance and the inducement of co-operation, albeit in a less brutal and more nuanced manner. Whether in prison or education, integrated power is realised through surveillance and extended and guaranteed through co-operation. And in both--and all such sites--"assessment" combines and produces both.

The growing use of assessment/punishment in higher educatio

Assessment has traditionally been a defining characteristic of the academic professional: assessment of students, and--generally to a lesser degree--of fellow professionals through peer review. A new, more exaggerated form of assessment has, however, become prevalent in recent years. The trend toward the "publish or perish" mentality has brought with it a new, rigid, punitive and hierarchical approach to assessment. The Research Assessment Exercise is merely one, albeit extreme, example of this tendency.

The Research Assessment Exercise (RAE): A Descriptive Introduction

The RAE was established by the Conservative government in June 1992 to accommodate its "wish to see selectivity in the allocation of research resources based on assessments of the quality of research" in higher education (Note 2). (HEFCE, 1994, para 4.) The first RAE was concluded in March 1994, publishing its findings in the form of a "league table" later that year. Our focus is on the second RAE, which refined the workings of the assessment structure and concluded its work on 31 March 1996, the results of which were made public in December 1996.

The RAE accords a ranking to every "unit of assessment" (UOA--most usually an academic department--in the United Kingdom). This ranking can be expected to exert a decisive influence over research funding allocations. "Units" are marked on a scale from 1 to 5 (with a new 5* category for star performers). A ranking of 3 is generally understood to be the minimum accepted standard necessary to warrant continued institutional support. However, "3" is now divided into 3A and 3B, with 3B likely to be judged the wrong side of the divide. Definitions are duly provided in the voluminous documentation accompanying the exercise--along with definitions for all of the many key words and terms employed. These definitions serve not only to clarify and guide, but limit and confine, participation and response.

The RAE is a massive operation, dominating the operation and orientation of higher education. No activity can take place without reference to it. The activity required to set up the Exercise was itself intensive. In 1994, 60 assessment panels were established to consider the submissions from 69 subject areas. Under direction from the funding councils, the Chairs--appointed by the funding bodies, on the advice of the 1992 RAE panel Chairs--were charged with assembling their teams, achieving a specified optimum degree of continuity in personnel (33%) with previous panels. Personnel selection was required to be based on detailed criteria including the "research experience of nominees and their standing in the research community" and "the need to secure representation from the research commissioning and user communities within commerce, industry, government and the public sector." (HEFCE, 1995, para. 4a) Chairs' recommendations for personnel would then require approval by the Chief Executive of the relevant funding body. (Ibid., para. 5a)

In line with the increasingly utilitarian re-assessment of research in higher education, and "in the light of the emphasis on developing the partnership between higher education and the users of research," some 1,000 invitations were issued, by the Chief Executive of the Higher Education Funding Council of England (HEFCE) on behalf of the four British funding bodies, to industrial and business and professional organisations for nominations for membership to the panel. (HEFCE, 1994, para. 18) Each panel was provided by the funding council with a Secretary. The Secretaries were responsible for ensuring that the elaborate procedures and regulations of the Exercise were carried out. One important regulation was that panels were "instructed to channel requests for clarification of data through the funding bodies and not to contact institutions directly." Similarly, any feedback they wished to give UOAs at the end of the process would also be channelled through the funding bodies. (HEFCE, 1995a, para 29)

Panels were charged with drawing up the assessment criteria for their own areas, taking into account "previous statements on the framework of the Exercise; advice from the funding bodies on policy and administrative considerations, and representations made by subject associations and other interested parties." (HEFCE, 1995b, para 4) Despite the appearance of a degree of freedom in establishing the criteria, it is important to recognize that a definition of "research" is provided by the funding councils which can not be challenged. The task of the panels is thus essentially to interpret this mandatory definition--"fine-tune" it to the specific requirement of the subject under review.

The common definition of Research reads:

> "'Research' for the purpose of the RAE is to be understood as original investigation undertaken in order to gain knowledge and understanding. It includes work of direct relevance to the needs of commerce and industry, as well as to the public and voluntary sectors; scholarship*; the invention and generation of ideas, images, performances and artefacts including design, where these lead to new or substantially improved insights; and the use of existing knowledge in experimental development to produce new or substantially improved materials, devices, products and processes, including design and construction. It excludes routine testing and analysis of materials, components and processes, e.g. for the maintenance of national standards, as distinct from the development of new analytical techniques.
>
> * Scholarship embraces a spectrum of activities including the development of teaching methods; the latter is excluded from the RAE." (HEFCE, 1995a, annex a)

The definition vaunts, above all else, the wider, specific benefits of the results of applied research to society and the economy. This seemingly indisputable and worthy objective can act to constrain criticism of social and economic values and norms. It valorises research as a means of production: research as production-line. It also implicitly suggests that the quality of such research is likely to benefit from intense processes of assessment and judgement.

One reason why scholarship such as development of teaching materials is excluded from consideration may be that it does not lead, concretely enough or quickly enough, to "ascertainable" benefits, commercial or otherwise. Teaching is doubtless not seen as in opposition or contradiction to such utilitarianism: its utilitarianism is merely of a longer-term kind, beyond the horizon avidly scanned by the RAE. That is, teaching, like research, is still a production-line, but one producing--moulding and integrating--workers (researchers) rather than products.(Note 3)

The decision to exclude such scholarship understandably proved controversial within the profession--"caused some difficulty," in RAE-speak. (HEFCE, 1995, annex a, para 19) Many academics see teaching-preparation as a legitimate contribution to, and an integral component of, their research activity. For the RAE, this legitimacy is conferred only when it "can be shown to embody research outcomes within the RAE definition." Such "embodiment"-- the production of appropriate, i.e. published, assessable output--precludes consideration of what has hitherto generally been regarded, and valued, as creative, original research. This research--a great body of work and output--is now "disembodied," relegated somehow to the status of "phantom" research; an incomplete production of thought. This view is simply dismissed--for reasons not explained--in the RAE: "the broader argument that the preparation of teaching material, as a form of scholarship, must generally be accepted as a research activity within the RAE is not accepted." (Ibid.) Such a blunt refusal starkly illustrates the arbitrary power of the Exercise. The nature, functions and effects of this power are those of an integrated disciplinary system. Understanding the workings of such a system can therefore illuminate the deeper implications of a process such as the RAE.

The Research Assessment Exercise: Operation and Effects

As mentioned, assessment of academic performance and "quality" in higher education has traditionally consisted of peer-review exercises operating within a hierarchical framework. While both features are retained within the RAE, hierarchical aspects take precedence, controlling and constraining the peer-review dimension. Likewise, the RAE is constrained by its location within broader hierarchical relationships. At the top of the hierarchy is the government, making pronouncements and establishing the mandate under which the funding councils must operate. The funding bodies dictate to the RAE panels they have approved. Once the rules of the process are established, the UOAs are obliged to reach the targets set for them by the panels. Ultimately, pressure is exerted on the individual academic, whose "output" and performance becomes bound to, and binds, all the links in this long chain of command. As a consequence, the existing hierarchical nature of the UOAs themselves becomes exaggerated.

Maintaining and monitoring such an elaborate hierarchy requires considerable levels of both surveillance and consent. Cooperation is vital at each level, as is "assessment," i.e. surveillance, of its effectiveness. A dynamic is established which serves to integrate and service the system. A "network" of power-relations between and within each level is produced, and continually reproduced, on the basis of the integration of those apparent polarities, surveillance and cooperation: "for although," as Foucault says, "surveillance rests on individuals, its functioning is that of a network of regulations from top to bottom, but also to a certain extent from bottom to top and laterally; this network "holds" the whole together and traverses it in its entirety with effects of power that derive from one another: supervisors perpetually supervised." (Foucault, 1977: 176-7)

The Integrated System and Surveillance

Foucault identifies and discusses "five distinct operations" at work within the integrated regime of disciplinary power. (Ibid., 182-3) How are they at work within the RAE?

The first operation "refers individual actions to a whole that is at once a field of comparison, a space of differentiation, and the principle of a rule to be followed." In the RAE, all individual research action is referred to its value as determined by the Assessors. This determination is made on the basis of standards of comparison and differentiation ostensibly set out for all to see, but actually open to a few--the Panels--to interpret. In the Exercise, while individual performance is assessed, it is the UOA concerned that is judged. For the RAE, the UOA is the "individual": UOAs are compared to each other as if they were Supra-Researchers. For the individual, the UOA is one of two "wholes," the other being The Exercise. Additionally, researchers within UOAs are compared to each other, and in many cases penalised or rewarded for success or failure in meeting goals set within the hierarchical structure of both the RAE and UOA.

As mentioned, latitude of interpretation--freedom of manoeuvre--for the assessors is built into the process. This latitude serves to constrain the freedom of manoeuvre for the assessed, the researcher, who is compelled to refer her or his output to injunctions which are both imperious and imprecise--indeed, whose imperiousness is enhanced and characterised by its very imprecision.

This characteristic is justified, indeed vaunted, in The Exercise thus: "The assessment process is not a mechanistic one." This claim gives Panels the right to remain vague in their pronouncements of what specifically is taken into consideration in arriving at a judgement. The individual researcher is told, for example, that publishing in "prestigious journals" or chairing "key conferences" will enhance their UOA's standing, though no definitive list is provided. Naturally, were such a list to be provided, it would be controversial and rightly condemned for its dictatorial audacity. The point being made is that this vagueness is not the consequence of wishing to accommodate others, but an essential mechanism in the accommodation and consolidation of the ultimately arbitrary power and remit of the assessors to assess.

The second operation is the differentiation of "individuals from one another, in terms of the following overall rule: that the rule be made to function as a minimum threshold, as an average to be respected, or as an optimum towards which one must move." In the Exercise, this function is exemplified by the designated number of publications sought. Following on from the results of the 1992 RAE ranking, the UOAs began to set targets for its individual researchers to aim for four publications in the four years between the Exercises. "Getting your four" became the mantra to "guide" the British academic. Here again a deliberately vague set of criteria left both the individual and the UOA struggling with a variety of unknowns: such things as the value (defined by the assessors) of different kinds of publications (books, chapters in books, articles in journals, etc), and the source of publication. Even the "four" is a variable. In an effort to avoid the appearance of a strictly quantitative approach, the funding councils have specified that the assessment panels should take into consideration "particular professional circumstances likely to lead to a reduced publication rate." (HEFCE, 1995b, para 12) In such cases, it is incumbent on the UOA to provide evidence of long term research projects or mitigating circumstances; extreme care must be taken if utilizing this clause lest the assessment panel fails to regard the reduced "output" as justifiable.

The third operation is one which "measures in quantitative terms and hierarchizes in terms of value the abilities, the level, the 'nature' of individuals." Here again we see the UOA doing to the individual what the RAE does to it. In accepting the "logic" put forward by The Exercise that certain forms of research "output" are superior to others, the UOA demands of its individual researchers that effort be made to attain certain standards. A single-authored book--quality is not an important consideration, or rather is assumed to be present in any such work--is given precedence and scores most heavily; more heavily, say, than a single-authored paper, which, if certain conditions are met, counts for more than a chapter in a book, and so on. The hierarchy of "good researchers" is thus established within the UOA, with--paradoxically--less value accorded to individual, thoughtful, long-term research.

Changing thus the "nature" of research changes also the "nature" of researchers. This is a prime example of what Foucault describes in the context of overtly penal institutions as assessment being far more concerned with the creation, through a system of rewards and punishments, of a certain type of individual than with the reform or improvement of individuals. In this case, the British government hopes to manufacture a "new breed" of researcher, more concerned both with their own hierarchical positioning and with the market-value of their research; market-value in terms of utility for "users," in business and elsewhere, and in terms of the contribution to the UOA's ranking, and thus its positioning within the UOA hierarchy. And this is so important because of its relation to funding and--as may well be seen increasingly in the near future--survival. This leads on to Foucault's fourth and fifth operations, which we consider together.

The fourth operation "introduces...the constraint of a conformity that must be achieved"; the fifth "traces the limit that will define difference in relation to all other differences, the external frontier of the abnormal." These two operations form the crux of the penal mechanism of the RAE. Put starkly but accurately, for the UOA non-participation can mean extinction. Equally, participation demands conformity to an array of specifications and definitions, all of which demarcate the normal from the abnormal, success from failure. Most importantly, the 3 ranking is widely regarded as marking the "external frontier," on one side of which lies "safety" in the sense of the continuation of probable adequate funding (at least until the next assessment).

The emphasis in these operations on conformity--normality--points to an apparently contradictory aspect of the disciplinary system much referred-to by Foucault: namely, that its concentration on the "difference" between individuals--their examination, assessment and consequent categorisation--is actually an insistence on a sameness, a uniformity and conformity. In the penal setting, for example, what matters is not that there are different types of individuals in prison, but that all individuals become--are reduced to being--different types of prisoner. However, in all disciplinary systems--and for both the individual researcher and the UOA--the following applies: "The perpetual penalty that traverses all points and supervises every instant in the disciplinary institution compares, differentiates, hierarchizes, homogenizes, excludes. In short, it normalizes." (Foucault, 1977: 183)

The Integrated System and Cooperation

For the integrated system to succeed, Foucault argues, cooperation is a necessary accompaniment to surveillance. It is also a consequence of it: the individual cooperates in part because s/he knows s/he is under surveillance. Thus, "good," cooperative behaviour has every likelihood of being rewarded; "bad," uncooperative behaviour, of being penalised. Cooperation also entails self-surveillance--one checks to ensure one is adequately mapping an entire research performance and planning to the requirements of those who will assess it--and the surveillance of fellow professionals: after all, a "non- cooperator," or under-performer, is capable of inflicting potentially calamitous damage on her/his colleagues.

Further, cooperation acts to endorse and legitimise the process of assessment and surveillance, and thereby the disciplinary system, as an integrated whole. This acts to fragment units which were previously cohesive. A dramatic example of this is the emergence of a "transfer market," as Departments (at least those who can, or make sacrifices to be able to) buy-up "star" players to strengthen their team and thus bolster their chances of "promotion" to a higher "league," in this case ranking. As in sport, such promotion brings with it increased money with which further valuable acquisitions can be made. As the journal Managing HE recently observed:

> "There has been no formal quantification yet of the transfer market, but analogies with football were reinforced by one contributor to a File on Four [BBC Radio] programme, who had pursued phone-calls at midnight and meetings in motorway service stations. ...one university 'losing' a professor has sent the new employer a bill for £0.5 million for intellectual property transfer in relation to a vital database developed for research.'" (Note 4) (Managing HE, 1996: 4)

University solidarities are put under further pressure as "high-ranking" Departments (who may have bought their rank in the above manner) look askance at those beneath them, fearful that they will be tainted by association and/or that they will be asked to subsidise them. Within Departments, collegial solidarities are undermined as researchers who may not meet targets set by the UOA are classified as "weak-links"; at which point, penalties against them may be exacted. Examples might be the imposition of heavier teaching and administrative loads, and the loss of research allowances, both financial and temporal (i.e. sabbaticals). Thus, features of the job once regarded as standard and unexceptional have been drawn into and deployed as part of an all-encompassing system of rewards and punishments designed to maximise "cooperation" with The Exercise.

The examples given above, and others like them, are ways in which both UOAs and researchers participate in and cooperate with the RAE not reluctantly but imaginatively, aggressively and competitively; such methods are overtly mandated, required or encouraged by The Assessors. This phenomenon begs questions about the "nature" of the profession itself and its ability to resist such an apparently clear threat to it.

It is perhaps easiest to understand the profession's complicity in its own surveillance and oppression by utilising the Gramscian notion of "spontaneous consent," a concept akin to the cooperative dimension of the integrated disciplinary system. In the case of The Exercise, consent is spontaneous--comes "naturally"--to the academic as a consequence of many years of systematic moulding of the professional personality. This moulding begins before entry into the profession, through the long years of being examined, assessed and rewarded as a student. Further examinations await, but the key test now is the professional's ability to examine, assess and either reward or punish. This ability is not only exercised on students but on colleagues in the culture of peer-review. The positive internalisation of this way of proceeding and being leads to the unquestioned--"spontaneous"--acceptance of disciplinary power, the ritualisation of which is the examination; a glorification of which is the RAE. Foucault sums up the nature and effects of this ritualisation thus:

> "The examination combines the techniques of an observing hierarchy and those of a normalizing judgment. It is a normalizing gaze, a surveillance that makes it possible to qualify, to classify, and to punish. It establishes over individuals a visibility through which one differentiates them and judges them. That is why, in all the mechanisms of discipline, the examination is highly ritualised. In it are combined the ceremony of power and the form of experiment, the deployment of force and the establishment of truth. At the heart of the procedures of discipline, it manifests the subjection of those who are perceived as objects and the objectification of those who are subjected." (Foucault, 1977: 184-5)

The RAE is a drastic imposition of such processes on a profession which itself practices them continuously and unquestioningly. This is not to say, however, that it is not changing that profession in important and disturbing ways. On the nature of research, on the value of teaching, and on the experience of students, the impact is proving profound.

Conclusions: The Impact of the RAE on Higher Education

We have argued that the RAE is an example of an exercise in disciplinary power as that term was understood by Foucault. The effect of the RAE on the academic profession can be seen on many levels, ranging from day- to-day stress and workload to the likely long-term nature and value of research and teaching. The RAE was intended to have a dramatic effect: it has had.

Impact on Research

The RAE represents a new phase in the "commodification" of academic research. Academics have long been expected to view publications as "assets," or what Ronald Barnett refers to as "academic currency" to trade in the pursuit of advancement. (Barnett, 1992) The RAE has linked commodification directly to the overall goal of making the intellectual community "competitive," with Departments adding up their members' currency in order to compete for declining government funding. The government's intention is to create a leaner, fitter, research "sector," providing in the words of the then Education Secretary "a national resource of knowledge and expertise for the benefit of our international competitiveness..." (Department for Education and Employment, 1996).

While the RAE claims not to be concerned with the number of publications, its imperative has encouraged researchers to publish more often in order to meet the pressure within their UOA to be "research active." On the face of it, the emphasis in the RAE on research productivity, and researchers as producers, is having the desired effect: more academics appear to be publishing with greater frequency. Producing more articles, however, is not the same as doing more research. The regurgitation and multiple-placing of articles is on the increase. This process, although intellectually untaxing, is time-consuming, reducing time and energy available for both fresh research and course review. Moreover, as more is being published, recent studies suggest that less is being read. (Daly, 1994)

The pressure on the UOAs is, as we have shown, subsequently placed on the individuals within. Indeed, in many cases, a system is instituted within the UOA which closely mirrors that of the RAE. Because so much is as stake, the individual academic must conform to dictates and her or his own research plans fall victim to larger forces. This necessarily begs larger questions about academic freedom. Individual researchers are coming under increasing pressure not to undertake complex and/or radical work which may not be able to be compressed into the Exercise's four-year cycle. Furthermore, pressure is exerted to produce work in specified forms and places, regardless of their appropriateness in the perhaps grander scheme of the individual's research project or career. Researchers who resist may hinder their own position within the UOA, potentially also jeopardising the UOA's ranking. Not only may this set of pressures act to undermine the position, confidence and job- satisfaction of researchers, it may act to prevent researchers with ambitious, long-term research programmes from either moving to other Departments, or even being hired in the first instance. It is more important than ever to be, and to be seen to be, a "safe bet."

Impact on Teaching and Learning

The RAE has had a negative impact on teaching in a number of ways, both ideological and practical. Ideologically, the previously cited definition of "research" used in the Exercise has both exacerbated the false split between teaching and research and tended to stress the superior relevance--in terms of the quality of intellectual endeavour, the practical benefit to the economic well- being of the country, etc.--of the latter. Practically, there has simply been less time for academics, scrambling for their "4," to devote to teaching.

In terms of both logistical and ideological emphasis, the Exercise has acted to devalue previously rewarding and esteemed aspects of the academic profession. Perhaps the most tangible example is the devaluation of course- and lecture-preparation, which, if done properly, involves extensive, high-quality research. It is, however, difficult to gauge the extent of the subsequent decline in course standards. Moreover, a crucial aspect of such preparation is interactive research; that is, reflecting on and incorporating the responses and attitudes of students. In other words, the Exercise not only reduces the learning experience of the student but the learning experience of the academic.

There are a whole range of ways in which the Exercise has tended to devalue both the academic's role as teacher and learner, and the student's intellectual, and ultimately personal, respectability and dignity. One major consequence is the diminished amount of time made available for out-of-class intellectual engagement with students--again, a loss to each party; a precluded interactivity. More directly, perhaps, time for meetings with students on strictly course-related work--discussion of draft essays, supervision meetings, etc.--is also reduced or made more pressurised. In broader terms, when courses are modified and adapted, academics will be highly unlikely to accept improvements from the students point of view that could reduce the amount of precious research-time they need to set aside to complete the research programme--their own, increasingly time-consuming course work. Other equally negative developments could be cited, but our purpose here is not to list them but, rather, consider the response to them by the academic profession.

Resisting the RAE

Resistance to the RAE is inevitable: it has been an abrupt and draconian intrusion into the profession, increasing the job-insecurity, and diminishing the job-satisfaction, of many academics. Despite its professed dedication to improving research quality, the Exercise is clearly a politically motivated prelude to closures and redundancies--an exercise in justifying, ultimately in the name of British economic competitiveness, a further fierce attack on the higher education sector. However, the resistance of academics has not translated into any noticeable degree of effective action. No doubt this in part because of the very fear the Exercise has generated. There may be a deeper explanation, however, suggested by our use of Foucault's concept of the integrated disciplinary system.

A Foucaultian analysis, such as that proffered above, suggests that higher education--indeed, education generally--is a Research Assessment Exercise: a competitive system run by--and, at least traditionally, operating to the powerful benefit of--disciplinarian technicians of reward and punishment. In such a system, it is precisely the combination of complicity and coercion which is integrative, irresistible, seemingly inevitable. For the academic the RAE is the equivalent of the examination, combining the negative and positive elements of the integrated system; the examination is a concentrated, spectacular exercise of surveillance, observation and normalisation. In this sense, the RAE is distorting the academic profession by taking its own logic and turning it against it.

A truly radical critique requires the contemplation of the desirability and necessity of a new, non-disciplinary logic. Such critiques have been offered in the past and, significantly, they have been articulated as part of a wider critique of society. (Freire, 1972; Dale, 1976; Illich, 1971; Gramsci, 1971). As a starting point, we might return to these earlier debates to renew our aquaintance with the ways in which education functions as part of the larger structures--political, economic and social--of discipline within society. It is only then that we will be able to make connections between our own actions as academics in disciplinary structures and the disciplinary structures to which we are subjected.

Notes

1. The authors' names have been placed alphabetically; the order does not denote an unequal contribution to the research and writing of this paper. The authors would like to thank Gavin Beckett and Jeannie Grussendorf for their helpful comments on an earlier draft of this paper.
2. UOAs are now coming under pressure to elaborate Five Year Research Plans, integrating each individual's research programme into a single, "coherent" programme, thus increasing the compression on researchers from above.
3. As argued above, this power is itself exercised within the context of an integrated, hierarchical, disciplinary system. The assessment criteria established by each Panel is expected to faithfully reflect what are in effect injunctions from above (the funding bodies).
4. The introduction of a Teaching Assessment Exercise is said to offer a corrective for this tendency, but early indications suggest that this Exercise will not consider the needs of students any more than the RAE considers the full value of research.

References

Barnett, Ronald. (1996). Linking Teaching and Research: A Critical Inquiry. *Journal of Higher Education*, 63 (6).

Dale, Roger et al. (1976). *Schooling and Capitalism: A Sociological Reader*. London, Routledge and Kegan Paul.

Daly, William T. (1994). Teaching and Scholarship: Adapting American Higher Education to Hard Times. *Journal of Higher Education*, 65 (1).

Department of Education and Employment (1996). "Shephard announces committee of inquiry into higher education," Press Release 56/96, 19 February 1996.

Foucault, Michel (1977). *Discipline and Punish*. London. Penguin.

Freire, Paulo, (1972). *The Pedagogy of the Oppressed*. London, Penguin.

Gramsci, Antonio, (1971). *Selections from the Prison Notebooks*. [Edited and translated by Quentin Hoare and Geoffrey Nowell Smith] London, Lawrence and Wishart.

Higher Education Funding Councils (1994). 1996 Research Assessment Exercise. RAE96 1/94,

Higher Education Funding Councils(1995) "Membership of Assessment Panels." RAE 96 1/95.

Higher Education Funding Councils(1995a) "Guidance on Submissions." RAE 96 2/95.

Higher Education Funding Councils(1995b) "Criteria for Assessment." RAE 96 3/95.

Illich, Ivan, (1971). *Deschooling Society*. London, Marion Boyars

Managing HE - for decision-makers in Higher Education (1996), "These are indeed tense times," Spring.

Richards, Huw (1996). "Transfer market Block." *Times Higher Education Supplement*, 27 December, p. 1.

About the Authors

Lee-Anne Broadhead

Lee-Anne Broadhead is Lecturer at the Department of Peace

Studies, University of Bradford in the U.K.

L.A.Broadhead@Bradford.ac.uk

Sean Howard

Sean Howard is editor of *Disarmament Diplomacy*, and is on the staff of the Acronym Institute, an institute based in London working on disarmament, arms control and non-proliferation issues.

sean@gn.apc.org

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

John Covaleskie
Northern Michigan University

Alan Davis
University of Colorado, Denver

Mark E. Fetler
California Commission on Teacher Credentialing

Thomas F. Green
Syracuse University

Arlen Gullickson
Western Michigan University

Aimee Howley
Marshall University

William Hunter
University of Calgary

Daniel Kallós
Umeå University

Thomas Mauhs-Pugh
Rocky Mountain College

William McInerney
Purdue University

Les McLean
University of Toronto

Anne L. Pemberton
apembert@pen.k12.va.us

Richard C. Richardson
Arizona State University

Dennis Sayers
University of California at Davis

Michael Scriven
scriven@aol.com

Robert Stonehill
U.S. Department of Education

Greg Camilli
Rutgers University

Andrew Coulson
a_coulson@msn.com

Sherman Dorn
University of South Florida

Richard Garlikov
hmwkhelp@scott.net

Alison I. Griffith
York University

Ernest R. House
University of Colorado

Craig B. Howley
Appalachia Educational Laboratory

Richard M. Jaeger
University of North Carolina--Greensboro

Benjamin Levin
University of Manitoba

Dewayne Matthews
Western Interstate Commission for Higher Education

Mary P. McKeown
Arizona Board of Regents

Susan Bobbitt Nolen
University of Washington

Hugh G. Petrie
SUNY Buffalo

Anthony G. Rud Jr.
Purdue University

Jay D. Scribner
University of Texas at Austin

Robert E. Stake
University of Illinois--UC

Robert T. Stout
Arizona State University

---

archives | abstracts | editors | board | submit | comment | subscribe | search

EPAA
• enter the archives • browse the abstracts • the editors • the edit board
• submit article • submit commentary • search • subscribe
volume: • 1 • 2 • 3 • 4 • 5 • 6 • 7

This article has been retrieved 887 times since April 27, 1998.

Education Policy Analysis Archives

Volume 6 Number 9 **April 27, 1998** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal.
Editor: Gene V Glass Glass@ASU.EDU.
College of Education Arizona State University,Tempe AZ 85287-2411 Copyright 1998, the EDUCATION POLICY ANALYSIS ARCHIVES.Permission is hereby granted to copy any article provided that EDUCATION POLICY ANALYSIS ARCHIVES is credited and copies are not sold.

SOCRATES Invades Central Europe

Joseph Slowinski
Indiana University

Abstract

The objective of this article is to explore the current reality faced by higher education students in Central and Eastern Europe and to draw out the implications of this current reality for policy makers in the future. In the article, I explore the influence of transnational corporations' training programs on education as it currently pertains to Central and Eastern European higher education and employment. In addition, multinational corporate entities exercise influence on European Union policy through the role of lobby organizations and activities. I explore the influence of these practices on education with an emphasis on the emerging importance of Western language skills. In addition, I focus on the European Union and its efforts to expand into Central and Eastern Europe in order to provide a focal point for analysis.

Introduction

> "If there ever was an all-European house, it had an upstairs and a downstairs.... There was the industrialized West, and then there was another, underdeveloped Europe to provide meals and servants--raw materials, food and cheap migrant labour. ...Europe in 2018 will consist of a Western superstate whose floors are scrubbed by Romanians or Poles, and a periphery of

> beggarly Bantustans (Ascherson, 1988:12)
>
> The nineties are also witnessing the continuation of important social, cultural, economic and political developments that affect higher education. Prominent among them are the globalization of the economy, the decline of the welfare state, and the commodification of knowledge. Since the fall of the Berlin Wall in 1989, there has been a deepening of the shift from Keynesianism to neoliberalism, and with it a wave of privatization and an increasing presence of market dynamics in social exchanges."(Schugurensky, 1997, p. 1)

At no other time in our history has the global proliferation of consumer markets been so pervasive. Global corporations produce and disseminate products throughout the world with ever increasing speed and scale. Since the fall of communism, these international corporations have expanded their markets into the post-communist nations. With this expansion of economic markets ensues a simultaneous global influence in economic, cultural and political arenas. In regard to education, particularly higher education, transnational corporations influence curriculum choices through training practices as well as language utilization. As we approach the twenty-first century, more and more global corporations are utilizing centralized training practices at corporate universities while opting for English as the language of instruction. This usage serves to promote English as a global lingua franca.

At the regional level, Central and Eastern European nations have been actively seeking admission to the European Union. Bound by the 1993 Copenhagen European Council agreement, CEE nations who have applied for EU membership must be actively engaged in developing a functioning market economy. Consequently, these CEE nations must demonstrate the introduction of active measures to reduce state owned enterprise and create a free market system(Note 1). Due to these measures, the CEE region has been a fertile ground for international investment. Consequently, on a regional level, transnational corporations have been granted access to CEE markets due to the necessary adoption of neo-liberal economic practices facilitated by the World Bank and the European Union (Note 2). Consequently, these powerful corporate entities influence language policy due to employment opportunities in Western as well as Central and Eastern Europe (Note 3). Due to their centralized training practices, access to employment is dependent upon western language skills and knowledge. This serves to privilege those in the region who have English language skills; those who have acquired English language skills maintain an advantage in the economic and labor markets. Consequently, transnational corporations have begun to play a major role in the field of higher education due to employment access based on western linguistic skill.

Furthermore, the European Union (EU) exerts influence in the region due to the involvement in education and developmental assistance (Note 4). Recently, Romania and Hungary became eligible for participation in European Union Community programs in the field of education, training and youth: SOCRATES, Leonardo da Vinci and Youth for Europe programs. This will serve to expand EU support of education carried out through TEMPUS from 1990 - 1996 (Note 5). Since 1990, the European Union has instituted the TEMPUS program which has provided educational travel opportunities for faculty and students. An examination of TEMPUS program data indicates that an unequal flow, from east to west, of faculty and students has occurred; many more university faculty and students from the east are traveling to the west than vice versa. I contend that these unequal mobility flows demonstrate an advantage for those with western language skills; CEE university students and faculty endowed with western language skills are afforded the opportunity to travel to Western Europe. This advantage along with the promotion of English language for training among transnational corporations will serve to devalue Central and Eastern European intellectual work. Through the introduction of SOCRATES, mobility flows will further exacerbate an unequal East-West flow of students and faculty. Through this unequal flow, those with Western European language skills will gain economic privilege through increased job opportunities. Furthermore, these mobility flows will lead to a diminished output of scholarly work published in Eastern and Central European languages.

In this article I will explore the influence of transnational corporations and the European Union on Central and Eastern Europe. I contend that global trends facilitated by transnational corporate training is facilitating a de facto linguistic advantage for those who have acquired English language skills. Furthermore, since most transnational corporations maintain a strong lobbyist structure in Brussels, transnational corporations act more in the role of policy development in the European Union. Simultaneously, participation in TEMPUS has operated to privilege university students with western language skills. Further participation in ERASMUS will further this trend; those university students with French, German or English language skills will maintain a privileged position due to access to university mobility programs. EU involvement in the region paired with the international influence of transnational corporations will lead to an increased level of economic stratification in Central and Eastern Europe. Those who have western language skills, especially English, will be granted access to study and employment opportunities not afforded those without these linguistic skills.

Globalization: English as Corporate Lingua Franca

With the collapse of Soviet communism and subsequent opening to international capitalism, global corporations were given the opportunity to expand operations and markets into Central and Eastern Europe. During the initial phases of transition, those CEE nations which were oriented towards a market economy were rewarded with large amounts of foreign investment (Note 6). Consequently, the World Bank and other multilaterals provided funding for those nations which were actively engaged in attempts of free-market liberalization and privatization(Note 7). Due to a need for capital and required by multilateral conditionality, CEE nations have opened its doors to global commerce. Yet, global commerce is simultaneously English speaking; five hundred and sixty-six of the top one thousand corporations in the world are located in English speaking nations(Note 8). Due to the economic power of global commerce centered in English speaking nations, English has become a global lingua franca. With markets expanding into CEE nations, those with English language skills will be given a privileged opportunity in the labor market. Consequently, English will increase in value as expansion of transnational corporation continues into the east.

With this economic expansion comes a de facto influence on University curricula. Since University students in CEE nations will be granted employment opportunities with English language skills, a demand for English language instruction at the university as well as at other levels of the educational system has been realized(Note 9). Due to the economic influence of global commerce, the connection between transnational corporations and institutions of higher education continues to merge. For example, Mallampally (1997) provides the example of two internationally renown business universities (i.e., Institut pour l'enseignement des méthodes de direction de l'enterprise (IMEDE) and the International Management Institute (IMI) which were originally founded as corporate training centers for Nestlé and Alcan. Levels of university and corporate connections will be discussed in more detail later. Yet, the connections current exist and will continue to flourish as global commerce grows. Consequently, a hierarchal system of influence exists; labor market requirements (facilitated by TNCs) influence universities which provide the credentialing and cultural knowledge needed to advance in a global society. The power of economic capital works to define the system of higher educational institutions; due to economic influence, the practices of these corporations facilitates public demands which operate to change university practices and knowledge distribution (Bourdieu, 1973).

Table 1 illustrates the training strategies of the largest transnational corporations. From Table 1, large transnational corporations provide more technical training as well as on-the-job training than small to medium TNCs. Since large transnational corporations rely more on their own training systems, access to this training is critical for employment with these firms. Due the location of corporate headquarters, access to training is dependent upon western language skills. For example, since 566 of the top 1000 corporations are located in English speaking nations, English will more likely be the language of training. Consequently, what appears to be more important criteria in hiring is the language skills through which TNCs provide training for employees. Therefore, western language skills become more critical for employment in large TNCs.

Table 1. Training of employees by TNCs

Type of Training & Region	Small to Medium TNCs	Large TNCs
On-the-Job Training	% Providing Training	% Providing Training
South, East & South-East Asia	61%	75%
Latin America	60%	69%
All Developing Nations	61%	73%
Technical Training	% Providing Training	% Providing Training
South, East & South-East Asia	46%	71%
Latin America	35%	74%
All Developing Nations	44%	73%

Source. Mallampally (1997).

From table 1, it can be ascertained that large transnational corporations tend to train employees at centralized locations. For example, in 1993, Nestlé trained 1200 workers from over 60 various nations at its Rive-Reine training center (Mallampally, 1997). As ERT (1989:35) explains about Nestlé training, due to global expansion:

> a special group is prepared for an international career. An initial on-the-job training mixed with classroom seminars is offered during a 1½ - 2½ years. For this group it is especially essential to look after the company's interests as though they were one's own, through (1) mobility or the willingness and ability to move about both physically (i.e., from one geographical area to another) and socially, (2) adaptability both in geographical and intellectual terms, and (3) linguistic skill. The minimum requirement is for two languages, the preference being English, French and Spanish.

Similarly, large TNCs such as McDonald's as well as Anderson Consulting maintain their own universities which operate as training centers. Yet, in order to conduct training in these central locations, employees must utilize a lingua franca. Consequently, workers who are hired are required to have gained linguistic capital in French, English or German. At Airbus Industries, a collaboration between Spanish, English, German and French companies, workers communicate in English (World Press Review, 1997).

Global commerce is influencing training practices throughout the world while TNC's training policies are influencing language acquisition as well as access to employment opportunities. Consequently, those who possess Western European linguistic capi tal remain in a privileged position. Transnational corporations located in Western Europe or moving into Eastern and Central Europe are continually influencing language policy. Since the EU is expanding into CEE nations, CEE university students are driven to acquire Western European linguistic capital.

European Union: Conflation of the Regional and Global

Bound by the 1993 Copenhagen European Council agreement, CEE nations who have applied for EU membership must be actively engaged in developing a functioning market economy. In accordance with this EU mandate, the EU clearly views CEE nations as potential consumers of Western European products (Note 10). EU policy makers desire to increase the economic competitiveness of Western European corporations. "Further integration and enlargement will help rapidly growing income in Central and East European countries translate into a continuous rapid growth of the West-European export market" (European Commission, 1997a). With the European Commission keen on expanding economic markets into Eastern and Central Europe, EU policies are created in an effort to perpetuate free market liberalization and privatization in CEE nations which further contributes to the influence of global corporations.

Like all large political institutions, the European Union has an ancillary collection of lobbyist organizations working for private corporate interests. One of the largest is the European Round Table of Industrialists (ERT). ERT is comprised of some of the largest multinational companies in the world. Of the forty-six companies ERT represents, twelve are listed in the top one hundred corporations in the world. European business continues to a major play on the world's stage. In 1998, 139 of the top 500 and 290 of the top 1000 corporations were located in Western Europe. In 1990, 168 of the world's top 500 corporations were based in Western Europe (Ikeda, 1996). Table 2 demonstrates the international economic power of many of the ERT corporations.

Table 2. ERT Corporations' Economic Power

Corporation	International Corporate Ranking (Rank of 1000)
General Electric	1
Royal Dutch Shell	3
British Petroleum	21
Unilever	33
Nestle	38
British Telecom	45
Daimler Benz	58
Ericsson	71
Siemens	77
Bayer	91
Veba	94
B.A.T. Industries	96

Source. Business Week (1997).

With such economic influence, it would be prudent to explore the relationship between ERT and the EU. For example, "[t]he aim of the ERT is to strengthen Europe's economy and improve its global competitiveness" (ERT, 1998). In order to accomplish this objective, ERT makes contact biannually (i.e., every six months) with members of the government which currently holds the EU presidency. This is due to the change in the EU presidency each six months. During these meetings, ERT presents working papers, reports or position papers outlining their policy in regard to critical issues influencing their corporate markets. In addition, ERT operates at the national level with its members facilitating contact with country level governmental and parliament members.

Like all lobby organizations, the European Round Table of Industrialists is a policy dissemination body which attempts to steer EU policy decisions. For example, ERT released "Education for Europeans--Towards the Learning Society" in March 1995. Interestingly, a July 14, 1997 report on the European Council's decision to admit Hungary, Romania and the Czech Republic for participation in Community programs for education, training and youth makes reference to a White Paper entitled, "Teaching and Learning - Towards the Learning Society." According to the European Council, this paper "defines the priorities of an education which is capable of carrying out its traditional tasks while integrating the new economics, technological and, above all, human aspects" (European Parliament, 1997: 20).

In addition, in 1997, ERT published a report "Investing in Knowledge: The Integration of Technology in European Education." Later the same year, the European Commission released "Towards a Europe of Knowledge." ERT's report emphasizes an information society which learns through cooperation with corporations. This sentiment was echoed by the European Commission. In section three of the European Commission report, "The Parties Involved," economic partners are emphasized. "There must be a commitment to securing greater involvement of the business sector" (European Commission, 1997d, p. 7). Topics raised by ERT through their policy papers seems to have an influential effect on European Commission policy. This influence extends directly to institutions of higher education through collaborate efforts with universities.

European University--Corporate Connection

ERT maintains an Educational Policy unit through which policy papers and collaboration with European Higher Education is facilitated. For example, the ERT joined together with the Standing Conference of Rectors, Presidents and Vice-Chancellors of the European Universities (CRE) to conduct research, publish policy papers and facilitate policy which would benefit corporate interest. CRE represents 500 universities in thirty nations. Two examples of collaborative publications include the following: European Approaches to Lifelong Learning (1992); and Lifelong Learning: Developing Europe's Future Capability: The Role of Industry - University Cooperation (1991). Through these publications, ERT promotes an increased partnership with institutions of higher education. Table 3 provides an example of suggested university - corporate collaborations currently being utilized by ERT member corporations.

Table 3. University--Corporate Collaboration

Training Type	Description
Inhouse Training	Individual University faculty utilized at corporate headquarters
Tailor-made programs	Universities construct program for corporation
Joint Collaboration	Combine accredited company training with external university courses
Pick and Mix	Select courses at a variety of institutions
Publicly funded adult education	Open University
Self-study	Distance Education Supported with Technology

Source. European Round Table of Industrialists (1991).

ERT desires universities to serve its interest through a concerted effort to promote life long learning; training costs will be reduced if universities aid corporations in training. For example, the American Society for Training and Development (ASTD) reported that $25 billion was spent annual in order to train poorly educated graduates (Vaughn, 1997). In addition to global economic influence, the European Round

Table of Industrialist is attempting to drive educational policy making throughout Europe through its partnership with CRE and policy papers disseminated to the European Union at its biannual meeting with the EU presidency. Yet, the corporations represented by ERT influence university students as well as language policy in CEE nations through its global economic strength.

Corporate Influence in Central and Eastern Europe

A recent survey conducted by Universum (1997) illustrates how Polish, Hungarian and Czech University students view language skills and their relationship to economic opportunity. Fifty-seven percent of the respondents indicated that "the ability to speak foreign languages" was a critical skill perceived as necessary to realize success in career plans. A large percentage realize that access to increased levels of employment in the European market is dependent upon the possession of language skills. Economic incentives, realized through language ability, remain the impetus for travel to Western Europe for education and employment opportunities. On the other hand, the large number of Romanians and Hungarians without knowledge of Western languages face an uncertain and perceived unfair future. World Press Review (1997:8) provides insight into their feelings:

> Hardest to take for many non-English speakers is the way the global language has divided the world into haves and have-nots: opportunities for knowledge, jobs, and advancement may be open to English speakers and closed to others. Career ads in French newspapers published in Belgium are often used in English, because multinationals increasingly regard mastery of the language as a job requisite.

With such an emphasis on foreign language acquisition as a requisite for employment, transnational corporations are endowed with a de facto influence over CEE university students. Due to the economic prowess of TNCs, CEE students look to these multinationals as a mechanism for upward mobility. Of those surveyed about companies that they would ideally like to work for, only four domestic companies (i.e., two banks, one telecommunications company and one brewery) made the top 25 "wish list" in total of the three countries represented in the survey. Interestingly, in two nations these companies were the number one choice. Generally, university students in the region are looking toward the West for economic prosperity; thirty-six percent of the respondents want to work for a multinational. This desire for employment at multinational corporations leads to a de facto influence on language programming at Central and Eastern European universities. For example, students seeking upward mobility realize the need to acquire Western European language skills which in-turn creates a demand for these languages at CEE universities. Fifty-seven percent of the university students indicated that acquiring a command of foreign languages is important to current career success. In other words, these university students recognize that acquiring linguistic capital (i.e., modern languages of Western Europe such as

German, French and English) is critical to upward mobility in the emerging new Europe. In addition, the influence of western linguistic capital can also be seen in the number of students from post-communist nations studying in the United States; 4780 students came to the U.S.A. during the 1991-92 year to study at American universities as compared with 18,032 during 1995-96 year (Moffet, 1996). The influence of global commerce on language acquisition can be seen through this recognition. Consequently, transnational corporations (TNC) create linguistic value through their global economic power. Students are clearly aware of this fact and equate economic success with working for this influential transnational corporations. Table 4 illustrates the university respondents top ten selections for ideal employment.

Table 4. CEE University Students Dream Jobs.

Poland	Czech Republic	Hungary
IBM	Komenci Banka	MOL
BMW	IBM	IBM
OPEL	Microsoft	Unilever
Microsoft	SPT Telecom	Coca Cola
General Motors	Citibank	BMW
Philips	Coppers & Lybrand	Andersen Consulting
Siemens	Hewlett Packard	Nestle
Bank Handlowy	Arthur Andersen	Mercedes-Benz
Arthur Andersen	Siemens	Audi
Sony	BMW	Danone

Source: Universum International (1997)

Surely, the majority of these companies are familiar to the reader. Students chose large TNCs; TNCs that represent some of the most successful and wealthiest corporations in the world. With this recognition comes a simultaneous realization of the language of their corporate headquarters (i.e., English, French, German). Obtaining linguistic capital for these university students represents an economic incentive for upward mobility. In addition, these languages operate as a transnational corporate ling ua franca allowing communication between native speakers of various vernacular languages. The influence of English and other Western languages facilitates university policy and curriculum change as well as a simultaneous demand by university students to acquire these linguistic skills.

EU Involvement in Central and Eastern Europe Higher Education

Beginning September 1, 1997, Romania and Hungary became eligible for participation in European Union Community programs in the field of education, training and youth: SOCRATES, Leonardo da Vinci and Youth for Europe programs (Note 11). In a ddition, the Czech Republic has recently became eligible to participate and it is anticipated that other CEE countries (i.e., Slovakia, Poland) will become eligible in the near future.

The European Union (EU) has been actively involved in Central Europe since 1988 (Note 12). As CEE nations realized more freedom, the level of EU involvement and cooperation from the EU increased. On December 18, 1989, the EU created PHARE in an effort to provide financial assistance and advice to post-communist nations (Note 13). As part of the creation of PHARE was the development of TEMPUS (Note 14). TEMPUS has been the EU's primary developmental assistance program in the field of education for Central and Eastern Europe. TEMPUS was initially implemented in order to meet the education and training needs of Hungary and Poland in an effort to support the initiatives of Phare. Yet, the EU soon realized Member Countries could benefit from expanding aid to other nations in the CEE region. In 1990, the Ministers of Foreign Affairs of G24 nations extended financial assistance to Czechoslovakia, Bulgaria, the German Democratic Republic and Yugoslavia

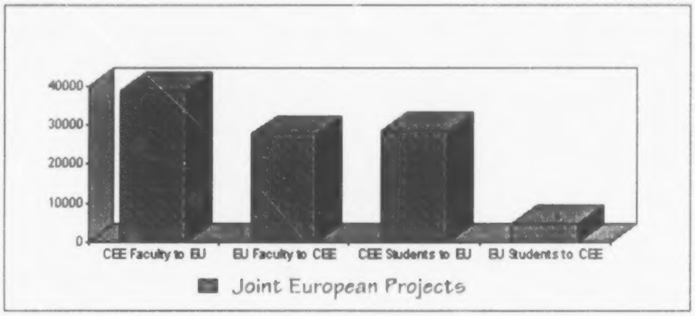
On May 7, 1990, the European Council created the TEMPUS program; Article 4 of the European Council decision outlines the objectives of the TEMPUS program (European Commission, 1997; European Commission, 1991). With the admission of associate countries into community programs, the EU will further its educational efforts first facilitated through TEMPUS (Trans European Cooperation Scheme for Higher Education). The objectives of TEMPUS are:

- to facilitate the coordination of the provision of assistance to the eligible countries in the field of exchange and mobility, particularly for university students and teachers, whether this assistance is provided by the Community, by its Member States or by third countries of the G24 group;
- to contribute to the improvement of training in the eligible countries, particularly in subject areas to which they give priority, and to encourage their cooperation, including joint cooperation, with partners in the Community, taking into account the need to ensure the widest possible participation of all regions of the Community in such actions;
- to increase opportunities for the teaching and learning in the eligible countries of those languages used in the Community and covered by the Lingua program and vice-versa;
- to enable students from the eligible countries to spend a specific period of study at university or to undertake industry placements within the Member States, while ensuring equality of opportunity for male and female students as regard participation in such mobility;
- to enable students from the Community to spend a similar period of study or placement in an eligible country;
- to promote increased exchanges and mobility of teaching staff and trainers as part of the cooperation process.

Of particular interest are the final two objectives: (1) to enable students to study or work in CEE nations; (2) to promote exchanges between EU and CEE faculty and students. Ideally, TEMPUS objectives promote cooperation as well as the exchange of ideas and cultures between citizens of EU and CEE nations. Yet, this is only true if a two-way flow of exchange occurs. Figures 1 and 2 illustrate faculty and student flows from 1990 to 1996.

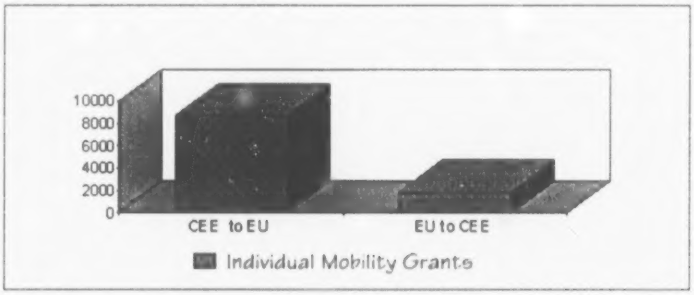
Consequently, an examination of the flow patterns of faculty and students to and from CEE nations should reveal if TEMPUS objectives have been met through exchanges and mobility flows. According to the European Training Foundation (1997a), large numbers of East-West and West-East exchanges have occurred. For example, 100,649 faculty and students have benefitted from TEMPUS funded Joint European Project exchanges as well as 10,624 individuals through Individual Mobility Grants (European Training Foundation, 1997b).

Figure 1. Faculty & Student Flows between EU and CEE countries (1990-1996)



Source. 1996 Tempus Yearbook.

Figure 2. Faculty & Student Individual Mobility Flows (1990-1996)



Source. 1996 Tempus Yearbook.

From this data, it becomes clear that an unequal flow of faculty and staff is occurring; faculty and students from CEE countries are traveling to the EU Member States in much greater numbers. What is particularly worrisome is the low number of students from EU member countries

BEST COPY AVAILABLE

traveling to CEE nations.

Since these students represent the future scholars from the region, the capability for access to knowledge from CEE nations is dependent upon translations from CEE nation students obtaining Western European language skills. Consequently, the potential for the loss or lack of distribution of academic work produced in CEE nations remains great due to the emphasis on Western European languages as well as these unequal academic mobility flows. Due to the emphasis on western languages, scholars will not have access to academic production written in CEE languages; those who will translate are more likely to translate from western to eastern languages. As English continues its global dominance, fewer scholarly journals will utilize CEE languages; the majority of the global intellectual products, journals and magazines, are published in a few languages: English, French, German, Spanish (Altbach, 1982)

Some may argue that this unidirectional flow is justifiable. After years under Soviet domination with strict regulations governing travel, faculty and staff from CEE nations desire travel opportunities to visit and explore Western Europe. In addition, universities in the region may not have the infrastructure capacity to support large numbers of west-east exchanges. Certainly, conversations with many students and colleagues from the CEE region indicate a desire to see Western Europe after oppressive communist policies. Yet, these mobility patterns are likely to result in a high degree of social-epistemological stratification based upon linguistic and other capital. For example, those who have acquired Western European language skills are able to participate in East-West exchanges. Current realities in Western European Academic exchange will only be exacerbated with the further entry of Eastern and Central scholars. For example, current participation of EU Member countries in staff mobility demonstrates the dependence on the dominant languages of Europe.

Enders' (1998) study of academic staff mobility in the European Union through ERASMUS demonstrates that a hierarchal value of languages exists in the European Union. Of all staff mobilities in 1990 - 91, English was utilized in 61 percent of course offerings where visiting faculty were lecturing at host institutions. French was the language of instruction in 27 percent of the courses, German 13 percent, Spanish 10 percent, Italian 9 percent and all other languages 2 percent. English is presently the academic lingua franca of choice.

In addition to the potential for decreased production of scholarly works in CEE languages, these opportunities created through university exchange could lead to future employment prospects with Western European universities and/or corporations leading to further economic stratification between university students in the CEE region who possess Western European language skills and those who don't. Consequently, economic stratification based on educational attainment will be further exacerbated by access given to those with western language skills. Access which provides upward mobility opportunities for those with these linguistic skills.

European Higher Education: Potential Impact on CEE Faculty and Students

In 1979, Jean-François Lyotard published *The Postmodern Condition: A Report on Knowledge*. His work acted as a catalyst for the postmodern movement but more importantly discussed his perceptions of the future of education and knowledge. Lyotard (1993, p. 4-5) writes:

> The relationship of the suppliers and users of knowledge to the knowledge they supply and use is now tending, and will increasingly tend, to assume the form already taken by the relationship of commodity producers and consumers to the commodities they produce and consume - that is, the form of value. Knowledge is and will be produced in order to be sold , it is and will be consumed in order to be valorized in a new production: in both cases, the goal is exchange. Knowledge ceases to be an end in itself, it loses its "use-value."

Due to this shift from knowledge as an end in itself to knowledge as symbolic capital, educational institutions must sell the acquisition of knowledge and skills to the consumer-student. Universities throughout the world are experiencing a shift from knowledge as an end in itself to knowledge as added value to the individual's symbolic captial. Consequently, universities are forging relationships with corporate entities as I have discussed previously in the paper.

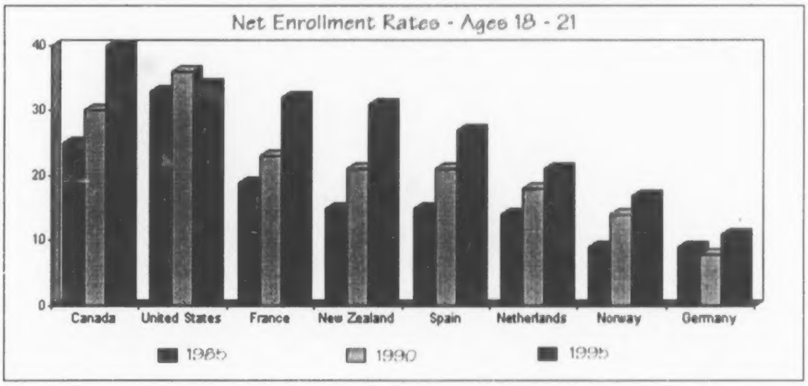
> Transformation in the nature of knowledge, then, could well have repercussions on the existing public powers, forcing them to reconsider relations (both de jure and de facto) with the large corporations, and more generally with civil service. (Lyotard, 1993, p. 6)

Lyotard's keen prognostication of the current reality we are witnessing in higher education is valuable for reflection on the impact of Central and Eastern Europe in regard to higher education. In CEE nations, the commodification of higher education is realized in two majors forms: privatization of education and the shifting profession of higher education.

With the elimination of communist state controlled education at the beginning of the 1990's came privitization efforts. International donor agencies including the European Union mandated privatization as a preresquisite to EU membership as well as receipt of donor aid. Consequently, the education sector was opened to private institutions offering all types of educational services. Through these privatization efforts, Central and Eastern Europe has witnessed the influx of hundreds (or perhaps thousands) of private institutions of higher education as well as universities headquartered in Europe or the United States with branch campuses in the region. Higher education has become a commodity. Since many of the programs strive to offer international models of business or authentic language instructors, local instructors are replaced with scholars who have knowledge of Western European business, law, economics models as well as the command of English or German. Could this lead to unemployment for those regional scholars without these languages or skills?

In addition to privatization efforts, higher education institutions in Europe as well as throughout the world are facing similar circumstances. Throughout the world in the past fifteen years, the percentage of students entering institutions of higher education as well as the percentage graduating has continued to rise substantially. Figure 3 demonstrates these trends in several OECD nations.

Figure 3. University Enrollment Rates (1985-1995)



Source. Peter (1997)

With large numbers of students entering higher education, universities are beginning to struggle financially. In addition, at the same time as large numbers of students are entering higher education, European governments are reducing per pupil funding. Consequently, a potential education crisis is being realized in Europe. To counter the economic reality of universal access to higher education, many nations have introduced increased levels of tuition and fees as well as stricter admissions policies (Schugurensky, 1998). Yet, with the opening of Central and Eastern Europe, western institutions may utilize the contacts brought about through TEMPUS and SOCRATES to fill needed lecture positions with CEE academicians.

Considering the influx of western scholars to Central and Eastern Europe as well as the access to university connections through mobility programs, Western European institutions of higher education potentially view Eastern European scholars as inexpensive labor to fill the needed positions to teach the masses entering higher education in Western Europe. Although these trends favor Western European institutions, Eastern scholars are eager for an income to support themselves and their families. Intellectual labor will, as Lyotard suggests, begin to resemble commodification. Advertisments may soon appear in LeMonde or Nepszabadsag written in English searching for lecturers to teach in Western Europe. But again we must be reminded of the linguistic capital

BEST COPY AVAILABLE

involved. English is a symbolic passport providing access to those who posess it.

Furthermore, this financial crisis may lead to less opportunities for Central and Eastern European students seeking placement in the universities with the best domestic reputation. Due to the financial crisis of higher education in the region, institutions of higher education are attempting to lure Western students who are willing to spend large sums of money on tutition to study there. With large numbers of foreign students coming to the region, university spaces at prestigious unviersities may begin to be filled with those who can afford to pay for tuition in hard currencies such as dollars, marks and pounds. Language continues to play a major role at the university level in the region.

Language as Neo-Colonial Agent or Catalyst for Upward Mobility?

Historically, the language of instruction at the university level or the acquisition of foreign languages has provided the recipient of educational services with prestige which could be exchanged as symbolic capital. For example, Latin as a dead language was the language of instruction at the higher education level and Hebrew, French and Greek were learned in order to train mental faculties. Acquisition of these languages offered potential wealth and prestige: similarly, those who acquire English today have access not afforded others who are educated to the same level. Perhaps we can draw parallels between what is occurring today with the past in regard that elite education has been instituted in an effort to restrict access. For example, elites (e.g., economically dominant individuals and families) attempted to restrict access through language of instruction. For example, Those today who don't have command of a western language are in a similar position; language skills can provide access or restrict it. Yet, there are fundamental differences as well.

In the past, Latin was used as symbolic capital when education was perceived as an end in itself. Elites differentiated themselves from others through knowledge acquisition and mental training rather than career aspiration. Since most students came from financial secure families, knowledge and mental training was the differentiating factor. They believed that a cultivated mind made an individual superior. Enlightment was driven by the pursuit of the rational and this would be achieved through education and mental training in part through the acquisition of foreign languages. Yet, today there has been a fundamental shift in the belief of the value of higher education. Today, a university education is seen more and more as a value added enterprise which provides the individual with symbolic capital to exchange in the labor market.

For example, Education adds value to human capital; higher education provides individuals with upward mobility opportunities. I would argue that those who obtained an education in Latin were also exercising an initiative to realize a superior economic position in exchange for their knowledge but this education was also sought more as an end in itself. Today's university student, particularly in Eastern and Central Europe, is driven much more by market forces. Since Central and Eastern Europe is realizing an economic crisis, these forces exercise great power on issues of university curricula and language of instruction.

University studies historically have operated as a mechanism to preserve elite status. Bourdieu (1973, 1990) illustrates this in French universities; most university students in French universities were children of the professional class. This process of elite preservation was experienced and continues today in African nations. After independence, elites maintained the colonial language due to the perception that this language afforded more opportunities for upward mobility. Yet, this was a condition of neo-coloialism where European language and cultural was perceived as superior. The people called for the language of instruction to remain the Western European language because they perceived these languages as leading to positions of economic superiority and prestige. Indeed, the French Cultural Ministry and the British Council are involved in maintaining the utilization of French and English in African nations and have established lucrative business operations similar to the establishment of western schools in Central and Eastern Europe. But, in Eastern and Central Europe, masses of people have rejected the colonial language of Russian and replaced it with English and German.

Indeed, as with Latin and the Colonial languages of French and English, language acquisition in CEE is driven by a desire for upward mobility. Language skills are viewed as a value added enterprise. I do not see this as a nation-state neo-colonial situation. But, multinational corporations are facilitating a neo-colonial condition where English is perceived as superior to the national languages of Central and Eastern European countries. Today, English is a desired product to be obtained due to the economic superiority of English speaking corporations. As I have demonstrated, the largest companies in the world are located in English speaking countries and utilize English as a working lingua franca. Language skills today are driven by economics. Transnational corporations are continuing to replace the nation-state as a global and regional force as people perceive employment opportunities with Western European companies or in Western European nations through labor mobility offered with admittance into the European Union. Those with language skills in Central and Eastern European nations are at a distinct advantage. Yet, how prolific is the possession of English or other Western European languages?

National Advantage: Western European Languages as Cultural Capital

Von Kopp (1996) suggests that in order to succeed in CEE nations during the transition phase, it is more important for citizens to acquire cultural capital than economic capital. For example, CEE students with foreign language skills are at a distinct advantage over students who have yet to acquire these skills. Yet, Hungary and Romania illustrate problematic realities in regard to linguistic capital. The Sixth Central and

Eastern Eurobarometer revealed that 79% of Hungarians and 78% of Romanians couldn't speak a second language well enough to converse in it (Note 15). In Hungary, similar results were obtained in earlier surveys; in 1979, 7% could speak a foreign language while in 1982, 13.9% surveyed could speak a foreign language (Radnai, 1994). Interestingly, the Czech Republic doesn't have such a language problem (Note 16). With such large numbers of citizens unable to converse in a second language, residents of Romania and Hungary are at a distinct disadvantage as the European Union expands. Yet, educational attainment in these nations plays a major role in access and privilege. As illustrated in Table 5, university educated Hungarians were twice as likely to have acquired language skills than secondary students. Educational attainment simultaneously brought about access to TEMPUS programs due to possession of western linguistic skill.

Table 5. Foreign Language Knowledge in Hungary

	< 8 Years of Primary School	8 Years of Primary School	Secondary School	College & University	Total Population
Knowledge of Foreign Languages	8.1	8.3	27.4	53.9	14.8
No Knowledge of Foreign Languages	91.1	91.7	72.6	46.1	85.2
# Sampled	3024	4359	1671	696	9750
Percent of Total	31.2%	44.7%	17.1%	7%	100%

Source. Radnai (1994).

With EU expansion comes mobility opportunities (Note 17). Yet, access to these mobility opportunities are dependent upon language skills. Those with English, German and French language skills have access to university studies in France, Germany and the United Kingdom as well as future employment opportunities with multinational corporations. Consequently, those university students with language skills can be considered to possess a form of symbolic capital: linguistic capital. Language skills are not capital in an economic sense but allow for symbolic exchanges which enable an individual an to extend the boundaries of her/his existence (Bernstein, 1973). A shift toward Western Europe has allowed those who possess Western linguistic and cultural capital to emerge as privileged. As we approach the twenty-first century, language skills conducive to communicating with American, Japanese and European Union member countries equate to job opportunities and economic advantage. This reality will further exacerbate the economic stratification brought about since transition due to an increased emphasis on credentialism; university graduates have a distinct economic advantage since the fall of Soviet communism (Note 18).

CEEPUS: Viable Solution or Contributor to Western European Linguistic Capital?

The Central European Exchange Program for University Students (CEEPUS) was initiated in January 1995. Realizing that "there had been substantial increases in 'East-West' exchanges, academic exchange among the new democracies had come almost to a complete halt" the Austrian Government facilitated a meeting of CEE ministers of higher education to discuss the creation of CEEPUS (Austrian Ministry of Education, 1995). Unlike TEMPUS, CEEPUS primarily operates as an East-East university exchange program. University exchanges occur between Hungary, Bulgaria, Romania, Poland, Slovakia, Austria, Croatia, Czech Republic and Slovenia institutions of higher education or departments.

Each nation pledges a number of scholarship months (see table 6); these months are considered CEEPUS currency. For example, this "currency" represents how many months institutions in each country will sponsor a student (i.e., host nations waive tuition costs for visiting students). Due to the age limit of the program (i.e., a maximum of 35 years of age), CEEPUS primarily promotes cultural and academic exchange between CEE students although university professors and graduate students are also encouraged to participate.

In regard to language, CEEPUS participation mandates the inclusion of courses in English, German and French. Consequently, these three languages operate as the official lingua francas of the CEEPUS program. In support of CEE languages is a type of course referred to as a "dual course." Dual courses consist of groups of participants from two countries where each group learns the language of the other group. Consequently, CEEPUS supports regional linguistic acquisition.

Although CEEPUS doesn't serve the quantity of TEMPUS mobility, as illustrated through the scholarship hours in table 6, the East-East nature of CEEPUS provides opportunities for CEE regional intellectual exchange and language learning. Consequently, institutions undergoing similar problems facilitated during post-communist transition can exchange ideas without the intervention of a western perspective. In addition, the East-East nature provides for a mechanism to preserve scholarly perspectives which can be endangered by global domination of academic journals written in Western languages.

Table 6. Scholarship Hours Pledged by Country

Nation	1995 - 1996	1996 - 1997
Austria	400	350
Bulgaria	100	100
Croatia	100	175
Czech Republic	-	100
Hungary	300	350
Poland	150	197
Slovakia	300	300
Slovenia	100	150

Source. Croatian CEEPUS Office (1997).

Is CEEPUS the only potential model for securing the language, culture and ideas of the region? Although I do believe that this is a best case scenario it is extremely limited. As can be ascertained through the number of available CEEPUS hours, this program is serving few students and faculty. Of course, there will be scholars from Central and Eastern Europe who will participate in academic mobility programs and return home to the region. These individuals are needed to be the disseminators of knowledge produced both in the West and the East. Without individuals who can provide access through translation in domestic academic scholarly journals to those without the linguistic capital to participate on the world academic stage, Central and Eastern European scholars may not be able to provide access to information in their own nations with the knowledge they generate. This has been common in developing nations where scholars supported by the Rockefeller and Ford Foundations publish in Western European language journals and do not publish in their own vernacular tongue.

Furthermore, the introduction of bilingual programs in Central and Eastern Europe may prove to provide access for those not able to take advantage of TEMPUS or SOCRATES mobility opportunities. These programs offer classes in the national language as well as through a western European language. Since 1990, there has an increase in the number of programs at the elementary, secondary and tertiary levels. Further research needs to be conducted to gauge the success of these domestic programs. These programs are built upon sanguine aspirations of those who believe that institutions of higher education can produce bi or trilingual students who can compete globally yet remain at home to serve the needs of the nation-state. Optimism is needed in a region where the entire educational system has been transformed in a short period of time.

Conclusion

I have attempted to raise some critical issues which have emerged in Eastern and Central Europe during post-communist transition. Globalization efforts by transnational corporations continue to promote the acquisition of English as an international lingua franca. Consequently, those who have acquired the English language maintain a de facto linguistic advantage. As I have demonstrated, the largest companies in the world are located in English speaking countries and utilize English as a working lingua franca. Language skills today are driven by economics. Transnational corporations are continuing to replace the nation-state as a global and regional force.

At the regional level, transnational corporations have emerged as policy makers; through influential lobbyist organizations, TNCs promote corporate self interest. Consequently, EU policies serve the interests of these powerful international entities.

An illustration of this influence can be seen through European Union expansion efforts into Central and Eastern Europe. EU expansion is driven by profit motivation and the expansion of commercialism to an additional 130 million CEE consumers. With this goal maintaining center stage, EU policy makers, encouraged by multinational lobbyists such as the ERT, will continue to develop educational policies which play a critical role in developing attitudinal, knowledge and skills conducive to consumer expansion.

At the national level, CEE students are flocking to Western European institutions through the TEMPUS program. Consequently, European Involvement in TEMPUS has facilitated a huge disparity between east and west faculty and student flows. I have argued that unequal flows from East to West, created through TEMPUS programming, have led to a privileged position for those who have acquired Western European language skills. Through TEMPUS mobility, these individuals gain access to employment and intellectual knowledge produced in Western Europe.

Furthermore, students from CEE nations continue to view Western Europe as the land of opportunity where those with the appropriate linguistic capital can reap economic gain and upward mobility. This dual process of CEE students desiring upward mobility and EU enculturation leads to a continual devaluation of CEE linguistic, cultural and academic arenas. If the current levels of economic stratification remain, students will continue to be driven toward English, French and German language programs. These policies will favor transnational corporations who will acquire the services of the brightest CEE university graduates with language skills leading to a CEE "brain drain".

Consequently, many CEE students are cashing in an obtaining Western cultural capital while rejecting their own legitimate culture. Unfortunately, those who don't obtain the necessary linguistic and social capital might become members of a working class periphery who wash the floors of those who exchanged their linguistic capital to the highest corporate bidder.

Notes

(1) The World Bank utilizes an Economic Liberalization index to indicate how economically liberal a nation in the CEE region has become. In

addition to the economic liberalization measure, the Heritage Foundation and the Wall Street Journal have created an Economic Freedom Index (Holmes et al, 1996). The liberalization index measures internal, external and entrance of new firms in the country (de Melo, 1997). Internal liberalization is weighted at .3 and is concerned with measuring domestic transactions such as abolition of state monopolies and price standardization. External liberalization is weighted at .3 and analyzes export controls such as tariffs and taxation. In essence, as is indicated by variables measured by the index, liberalization is a measure of free market practices. In order to receive funding from multilateral organizations such as the World Bank, CEE nations must strive for and maintain a high economic liberalization index since this index is referenced when determining developmental assistance.

(2) To receive developmental assistance money, CEE nations must be moving toward free market practices as well as reducing state owned properties and enterprise.

(3) International companies such as Ford or General Electric have either located operations or bought previously owned operations located in Central and Eastern Europe.

(4) Total aid funneled from the EU to the 12 CEEC nations between 1990 - 94 was 33.8 billion ECU; this represented 45 percent of all donor funding received in the region. In fact, the EU has become the fifth largest aid donor to the region . In 1995, the EU provided $7.1 billion which represented 10.5 percent of all donor aid to the region by OECD countries.

(5) Trans-European Cooperation Scheme for Higher Education

(6) Since the collapse of Soviet style communism in 1989, countries of Eastern and Central Europe have been inundated with donor assistance activities from the West. Foreign direct investment surpassed $46 billion in 1996 up 60% from the year before (HVG, 1997). Of this $46 billion, Hungary leads transition nations of Eastern and Central Europe with a cumulative $13.9 billion with Poland second at $9.1 billion; Hungary's foreign direct investment represents 30% of the region's total while Poland's 17%. Russia is third with a 13% while the Czech Republic maintains 12%. Hungary, the Czech Republic, Poland, the Slovak Republic and Slovenia together received approximately 70% of all foreign direct investment in the region.

(7) In an effort to examine the influence of a nation's neo-liberal economic status on EU donor funding, I ranked Phare countries by their 1990 World Bank economic liberalization index value and compared this with each nation's corresponding level of average annual Phare support from the EU(1990-1996). Since the level of Phare funding varied in regard to the number of years each nation has been funded, I calculated average annual levels of funding and utilized this value for comparison. Results of Spearman-Rho rank order correlational analysis revealed a significant positive correlation of = .86, $p < .005$ (.735). From this result, it would

indicate that multilateral funding offered by the European Union is contingent upon each nation's level of economic liberalization.

(8) Eight of the top 10, 67 of the top 100 and 566 of the top 1000 corporations in the world are headquartered in English speaking nations.

(9) The elimination of Russian language courses provided opportunities for English and German as well as other western languages: Spanish, French, etc. For instance, during the 1990 - 91 school year, three French, one Spanish and three German schools were opened in Czechoslovakia (Van Kopp, 1992). In Hungary, German is the most widely studied and used language (Radnai, 1994). In the 1989 - 90 school year, due to these transitional language policies, Poland experiencing a chronic need for more than 25000 English teachers but supply was limited to approximately 1500 teachers who were available to teach English (Vulliamy & Webb, 1996). Hungary, in order to meet the need of German and English language instructors, facilitated an extensive language retraining program. Russian teachers were trained in English, German or another foreign language (e.g., French or Spanish) to fill the demand for foreign language instruction generated from reform efforts.

(10) The European Commission believes that "[t]he economic effects of enlargement will undoubtedly be beneficial for the Union on the longer run. Enlargement will mean the creation of a larger economic area, with up to 500 million consumers, compared to the current 370 million....... Further integration and enlargement will help rapidly growing income in Central and East European countries translate into a continuous rapid growth of the West-European export market (Agenda 2000, 1997).

(11) Education in the European Union is governed by Article 126 and Article 127 of the Maastricht Treaty. SOCRATES is the European Community action program for cooperation in the field of education. SOCRATES was first adopted for EU member states on 14 March 1995 by Decision 819/95/CE. For an overview of the program visit http://europa.eu.int/en/comm/dg22/socrates.html. For information regarding SOCRATES expansion into Eastern and Central Europe visit http://europa.eu.int/en/comm/dg22/socrates/new-co.html

(12) European Union involvement in Eastern and Central Europe was facilitated by the signing of mutual agreements for cooperation on June 25, 1988 between COMECON (Council on Mutual Economic Assistance) and the European Community. This common declaration acted as the catalyst for developing cooperation between the EU and communist nations of Eastern and Central Europe. For example, Hungary signed agreements with the EC in September, 1988 as well as Poland who signed agreements in September 1989. These agreements were followed by the creation of PHARE (Phare (Pologne Hongrie Aide a la Reconstruction Economique) is the French word for lighthouse. Economic aid and advice from the European Union was intended to shine the light on the path back to Europe) on December 18, 1989.

(13) PHARE (Pologne Hongrie Aide a la Reconstruction Economique) is the French word for lighthouse. PHARE was created in order to shine a light (through financial aid and advice), for CEE nations, on the path back to Europe. During the 1990's, the European Union (EU) has contributed substantially to Central and Eastern Europe. From 1991-95, more than 60 percent of European Union donor aid for education went directly to the Central and Eastern European Countries (CEEC) and the Newly Independent States (NIS). From 1990 to 1995, the EU provided ECU 5.3 billion to Albania, Bulgaria, the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia and Slovenia.

(14) Since 1990,according to the European Training Foundation, which know operates TEMPUS, more than 90,000 staff and students from 7000 higher education institutions have benefitted from East-West and West-East exchanges. In addition, 3000 sets of lecture notes have been written as well as 15,000 courses have been created or updated. About 500 institutions have been founded or restructured, and roughly 2000 new books have been published.

(15) In November, 1995, the European Commission interviewed 20,278 residents of 19 Central and Eastern European nations. Countries surveyed include: Albania, Armenia, Belarus, Bulgaria, Croatia, Czech Republic, Estonia, Fyrom, Georgia, Hungary, Kazakhstan, Latvia, Lithuania, Poland, Romania, Russian Federation, Slovak Republic, Slovenia, Ukraine. This survey included all possible languages spoken (e.g., knowledge of Russian).

(16) The Central and Eastern Eurobarometer found that 36% of Czech citizens can speak Russian, 33% can speak German and 16% can speak English.

(17) As a citizen of the European Union an individual is supposedly free to live, work and set up business anywhere in a EU member country. See the abc of European Union - Citizenship at: http://europa.eu.int/abc-en.htm

(18) Prior to transition, due to centralized planning, school graduates in Central and Eastern European nations had many opportunities for employment regardless of educational attainment. Education was encouraged but was not as critical to social mobility as it is today. Those who have received a university degree have a substantial economic advantage today. Those with a minimal elementary education have witnessed the largest economic decline. For instance, 10% of the unemployed in the Czech Republic have only an elementary education (OECD, 1996a). In Poland, 20% of those with a vocational education are unemployed as well as 17.4% of those possessing only a basic education (OECD, 1996b). In fact, in Poland, only 4% of those with a higher education are unemployed. In the Czech Republic only 0.5% of those with a university degree are unemployed. Educational attainment has also influenced the quality of life. Furthermore, in Czechoslovakia in 1988, female University graduates earned 1.4 times that of a female with vocational training while the difference was 1.28 for men (UNDP, 1996).

In 1992, these values were 1.43 and 1.34 respectively. Possession of a university degree has become more and more lucrative in Central and Eastern Europe since the collapse of communism and centralized planning.

References

Altbach, P.G. (1982). Servitude of the mind? Education, dependency, and neocolonialism. In Arnove, R., Altbach, P. G., and Kelly, G.P. (Eds.), *Emergent issues in education.* Albany, New York: State University of New York Press.

APRODEV (May, 1997). PHARE and Civil Society. APRODEV Bulletin. [On-line]. Available at: http://carryon.oneworld.org/aprdev/may97_3.htm

APRODEV (December, 1994). The EU and Eastern Europe. APRODEV Bulletin. [On-line]. Available at: http://carryon.oneworld.org/aprdev/may97_3.htm

Ascherson, N. (1988, December 11). Below stairs in Europe's house. *The Observer*, 12.

Austrian Ministry of Education. (1995). *CEEPUS: Studying with friends - The Central European Exchange Program for University Students.* [On-line]. Available at: http://www.bmwf.gv.At/1bm/texts/95-5/ceep.htm

Bernstein, B. (1973). Social class, language and socialization. In Abramson, A. S. (Ed.). *Current trends in linguistics, 12.* Moulton.

Bourdieu, P. & Passeron, J. (1990). *Reproduction in education, society and culture*. London: Sage.

Bourdieu, P. (1973). Cultural reproduction and social reproduction. In Brown, R. (Ed.). *Knowledge, education, and cultural change*. (pp. 71 - 112). London: Tavistock.

Business Week. (July, 7 1997). *The Business Week global 1000.*

Clark, J. (1996). Developing competition policy in transition countries. *OECD Transition Brief, 4.* [On-line]. Available at: http://www.oecd.org/sge/ccet/trans4/competit.htm

Croatian CEEPUS. (1997). [On-line]. Available at: http://www.mzt.hr/mzt/hrv/medjunar/ ceepus/ceepus.htm

Dauderstddt, M. (1993). *The EC and Eastern Europe: The light is fading in the lighthouse.* Bonn, Germany: Freidrich Ebert Stiftung.

de Melo, M, Denizer, C. & Gelb, A. From plan to market: Patterns of transition. *Transitions, 12.* [On-line]. Available at:

http://www.worldbank.rog/html/prddr/trans/dec95/melo.htm

Enders, J. (1998). Academic staff mobility in the European Community: The Erasmus experience. *Comparative Education Review, 42*(1): 46 - 60.

European Commission. (1997a). *Agenda 2000: The effects on the Union's policies on enlargement to the applicant countries of Central and Eastern Europe*. [On-line]. Available at: http://europa.eu.int/comm/dg1a/agenda2000/en/impact

European Commission. (1997). *Eurobarometer, 46*. Brussels, Belgium: European Commission.

European Commission. (1997). *Tempus Yearbook 1997/98*. Brussels, Belgium: European Commission.

European Commission. (1996). *Central and Eastern Barometer, 6*. Brussels, Belgium: European Commission.

European Commission. (1996). *Eurobarometer, 45*. Brussels, Belgium: European Commission.

European Commission. (1991). *Tempus annual report*. Brussels, Belgium: European Commission.

European Parliament. (1997). *European Parliament Report: Legislative Proposal for a Community Decision regarding Hungary, Romania and the Czech Republic's involvement in Community Programs*. [On-line]. Available at: http://europarl.eu.int/dg1/a4/en/a4-97/a4-0248.htm

European Round Table of Industrialists. (1997). [On-line]. Available at: http://www.ert.be

European Round Table of Industrialists. (February, 1997). *Investing in knowledge: The integration of technology in European education*. Brussels, Belgium: European Round Table of Industrialists.

European Round Table of Industrialists. (February, 1995). *Education for Europeans: Towards the learning society*. Brussels, Belgium: European Round Table of Industrialists.

European Round Table of Industrialists. (June, 1992). *Lifelong learning: Developing Europe's future capability - The role of industry-university cooperation*. Brussels, Belgium: European Round Table of Industrialists.

European Round Table of Industrialists. (February, 1989). *Education and European competence*. Brussels, Belgium: European Round Table of Industrialists.

European Training Foundation. (1997a). [On-line]. Available at: http://www.eft.it

European Training Foundation. (1997b). Tempus annual report 1996: Phare & Tacis. [On-line]. Available at: http://www.

European Union (1996). *European Commission: The Phare Programme Annual Report 1995.* Brussels: European Union. Document Reference: P/EN/08.96/02/02/11/B

Holmes, K. R., Johnson, B.T., Kilpatrick, M. (1996). 1997 index of economic freedom. *Transitions, 12.* [On-line]. Available at: http://www.worldbank.rog/html/prddr/trans/ nd96/doc11.htm

Hungary CEEPUS. [On-line]. Available at: http://www.tpf.iif.hu/ceepus/

HVG (1997). Hazai is Vilag Gazdasag. [On-line]. Available at: http://www.hvg.hu

Ikeda, S. (1996). World Production. In Hopkins, T. & Wallerstein, I. (Eds.) *The age of transition: Trajectory of the World-System 1945-2025.* Leichhardt, NSW, Australia: Pluto Press.

Kniffin, K. (1997). Serving two masters: University presidents moonlighting on corporate boards. *Multinational Monitor, 18(11).* [On-line]. Available at:http://www.essential.org/monitor/ hyper/mm1197.05.html

Lyotard, J. (1993). *The postmodern condition: A report on knowledge.* (9th Ed.). Minneapolis: University of Minneapolis Press.

Mallampally, P. (1997). Transnational corporations and human resource development. *Prospects, 27* (1): 55 - 76.

Mateju, P. & and Rehakova, B. (1996). Education as a strategy for life success in the post-communist transformation: The case of the Czech Republic. *Comparative Education Review, 40*(2): 158 - 176.

Moffett, J. (1996). Eastern Europe/former U.S.S.R.: More students coming to study in U.S. *Radio Free Europe Radio Liberty.* [On-line]. Available at:http://www.rferl.org/nca/features/1996/12/F.RU.961204154150.htm

Organization for Economic Co-operation and Development (1996a). *Reviews of national policies for education: Czech Republic.* Washington, D.C.: Organization for Economic Co-operation and Development.

Organization for Economic Co-operation and Development (1996b). *Reviews of national policies for education: Poland.* Washington, D.C.: Organization for Economic Co-operation and Development.

Organization for Economic Co-operation and Development (1996c). *Secondary education systems in PHARE countries: Survey and project* *proposals*. Washington, D.C.: Organization for Economic Co-operation and Development.

Peter, D. (1997, October 4). Inside the knowledge factory. *The Economist*.

Radnai, Z. (1994). The educational effects of language policy. *Current Issues in Language and Society, 1* (1): 65 - 92.

Schugurensky, D. (1998). Higher education restructuring in the era of globalization towards a heteronomous model? In Arnove, R.F. & Torres, C. (Eds.). *Reframing comparative education: The dialectic of the global and local.* Boulder, Colorado: Rowman and Littlefield.

Slovak CEEPUS. (1997). [On-line]. Available at:http://www.kar.elf.stuba.sk/ceepus

Von Kopp, B. (1996). *Elite and education in the process of post-communist transformation - The case of the Czech society.* Paper presented at the World Comparative Education Society Biannual Conference. [On-line]. Available at: http://www.edfac.usyd.edu.au/projects/wcces96/ papers/vonkoppb.pdf

Von Kopp, B. (1992). The Eastern Europe revolution and education in Czechoslovakia. *Comparative Education Review, 36*(1): 101 - 113.

Vulliamy, G. & Webb, R. (1996). Education during political transition in Poland. *International Journal of Educational Development, 16*(2): 111 - 123.

United Nations Development Program. [On-line]. 1997 Human Development Index. Available at: http://www.undp.org

Universum International (1997). The Central European Study. [On-line]. Stockholm, Sweden: Universum International. Available at: http://www.universum.se/international/surveys/ces-96.

Vaughn, J. (1997). Big business and the blackboard: A winning combination for the classroom? *Journal of Law and Education, 26* (2): 35 - 46.

World Press Review. (1997). The world speaks English: Winning the language wars. *World Press Review, 44* (10): 6 - 8.

About the Author

Joseph Slowinski
Indiana University
4228 Education
Bloomington, Indiana 47405
Email: Joeslow@Indiana.edu

Joseph Slowinski currently is Associate Instructor at Indiana University where he teaches "Computers in Education." In addition, he is involved in post-communist education scholarship and serves as Assistant Editor of the Institute for the Study of Russian Education newsletter. He has earned a M.Ed. in Educational Administration and Supervision and is currently pursuing a Ph.D. in Educational Leadership and Policy Studies with a focus on International and Comparative Education. Over the course of his career he has taught in England, Hungary and Switzerland.

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

John Covaleskie
Northern Michigan University

Alan Davis
University of Colorado, Denver

Mark E. Fetler
California Commission on Teacher Credentialing

Thomas F. Green
Syracuse University

Arlen Gullickson
Western Michigan University

Aimee Howley
Marshall University

William Hunter
University of Calgary

Daniel Kallós
Umeå University

Thomas Mauhs-Pugh
Rocky Mountain College

William McInerney
Purdue University

Les McLean
University of Toronto

Anne L. Pemberton
apembert@pen.k12.va.us

Richard C. Richardson
Arizona State University

Dennis Sayers
University of California at Davis

Michael Scriven
scriven@aol.com

Robert Stonehill
U.S. Department of Education

Greg Camilli
Rutgers University

Andrew Coulson
a_coulson@msn.com

Sherman Dorn
University of South Florida

Richard Garlikov
hmwkhelp@scott.net

Alison I. Griffith
York University

Ernest R. House
University of Colorado

Craig B. Howley
Appalachia Educational Laboratory

Richard M. Jaeger
University of North Carolina--Greensboro

Benjamin Levin
University of Manitoba

Dewayne Matthews
Western Interstate Commission for Higher Education

Mary P. McKeown
Arizona Board of Regents

Susan Bobbitt Nolen
University of Washington

Hugh G. Petrie
SUNY Buffalo

Anthony G. Rud Jr.
Purdue University

Jay D. Scribner
University of Texas at Austin

Robert E. Stake
University of Illinois--UC

Robert T. Stout
Arizona State University

---

archives | abstracts | editors | board | submit | comment | subscribe | search

This article has been retrieved 1726 times since May 22, 1998.

Education Policy Analysis Archives

Volume 6 Number 10 **May 22, 1998** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal.
Editor: Gene V Glass Glass@ASU.EDU.
College of Education Arizona State University,Tempe AZ 85287-2411

Copyright 1998, the EDUCATION POLICY ANALYSIS ARCHIVES.Permission is hereby granted to copy any article provided that EDUCATION POLICY ANALYSIS ARCHIVES is credited and copies are not sold.

Educational Standards and the Problem of Error

Noel Wilson
School of Education
The Flinders University of South Australia

Abstract

This study is about the categorisation of people in educational settings. It is clearly positioned from the perspective of the person categorised, and is particularly concerned with the violations involved when the error components of such categorisations are made invisible.

Such categorisations are important. The study establishes the centrality of the measurement of educational standards to the production and control of the individual in society, and indicates the destabilising effect of doubts about the accuracy of such categorisations.

Educational measurement is based on the notion of error, yet both the literature and practice of educational assessment trivialises that error. The study examines in detail how this trivialisation and obfuscation is accomplished.

In particular the notion of validity is examined and is seen to be an advocacy for the examiner, for authority. The notion of invalidity has therefore been reconceptualised in a way that enables epistemological and ontological slides, and other contradictions and confusions to be highlighted, so that more genuine estimates of categorisation error might be specified.

Contents

- **Part 1: Positioning**
  - Chapter 1: Positioning the study: content and methodology
  - Chapter 2: Positioning the writer: experience
  - Chapter 3: Positioning the writer: philosophy and value
- **Part 2: Context**
  - Chapter 4: Power relations
  - Chapter 5: Power relations in educational settings
  - Chapter 6: Standards, myth and ideology
- **Part 3: Tools of analysis**
  - Chapter 7: Four frames of reference
  - Chapter 8: Equity, frames and hierarchy
  - Chapter 9: Instrumentation
  - Chapter 10: Comparability
  - Chapter 11: Rank orders and standards
  - Chapter 12: An inquiry into quality
- **Part 4: Error analysed**
  - Chapter 13: Four faces of error
  - Chapter 14: What do tests measure?
  - Chapter 15: The psychometric fudge
  - Chapter 16: Validity and reliability
- **Part 5: Synthesis**
  - Chapter 17: Error and the reconceptualising of invalidity
- **Part 6: Application**
  - Chapter 18: Competencies, the great pretender
  - Chapter 19: National tests and university grades
- **Part 7: Concluding statement**
  - Chapter 20: Out of the fog
- **References**

Acknowledgments

I wish to acknowledge the help of staff and students at the Flinders Instituite for the Study of Teaching for their help, support, encouragement, stimulation and companionship over the past three years.

In particular I want to acknowledge the support of my supervisor, John Smyth, for his courage for accepting me as a student in the first place, for his clear and incisive help when I asked for it, and sometimes when I didn't, and most importantly for showing me that there are still persons working in hierarchical systems who have been able to maintain their integrity in their search for truth and justice.

About the Author

Noel Wilson

Noel Wilson can be reached via Solveiga Smyth at
Solveiga@flinders.edu.au

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

John Covaleskie
Northern Michigan University

Alan Davis
University of Colorado, Denver

Mark E. Fetler
California Commission on Teacher Credentialing

Thomas F. Green
Syracuse University

Arlen Gullickson
Western Michigan University

Aimee Howley
Marshall University

William Hunter
University of Calgary

Daniel Kallós
Umeå University

Thomas Mauhs-Pugh
Green Mountain College

William McInerney
Purdue University

Les McLean
University of Toronto

Anne L. Pemberton
apembert@pen.k12.va.us

Richard C. Richardson
Arizona State University

Dennis Sayers
University of California at Davis

Michael Scriven
scriven@aol.com

Robert Stonehill
U.S. Department of Education

Greg Camilli
Rutgers University

Andrew Coulson
a_coulson@msn.com

Sherman Dorn
University of South Florida

Richard Garlikov
hmwkhelp@scott.net

Alison I. Griffith
York University

Ernest R. House
University of Colorado

Craig B. Howley
Appalachia Educational Laboratory

Richard M. Jaeger
University of North Carolina--Greensboro

Benjamin Levin
University of Manitoba

Dewayne Matthews
Western Interstate Commission for Higher Education

Mary P. McKeown
Arizona Board of Regents

Susan Bobbitt Nolen
University of Washington

Hugh G. Petrie
SUNY Buffalo

Anthony G. Rud Jr.
Purdue University

Jay D. Scribner
University of Texas at Austin

Robert E. Stake
University of Illinois--UC

Robert T. Stout
Arizona State University

---

archives | abstracts | editors | board | submit | comment | subscribe | search

Part 1: Positioning

Chapter 1: Positioning the study: content and methodology

Chapter 2: Positioning the writer: experience

Chapter 3: Positioning the writer: philosophy and value

Chapter 1: Positioning the study - content and methodology

Summary of the study

The project grew out of a general critique of assessment theory and practices, and in particular of the way in which the notion of error in measurement is obfuscated.

The fundamental research question that informed this study is:

> How is error in measurement of standards obscured in most practical events involving assessment of persons?

The study that subsequently developed

- Clearly positions the writer in terms of the experience, philosophy and values that he brings to this study.
- Develops some tools of analysis of the educational assessment process that enables a more stringent critique of the nature and extent of error in the measurement of standards.
- Establishes the centrality of the notion of the educational standard to the categorisation, production and control of the individual in society.
- Shows how the professional literature on educational measurement is based on the notion of error, and at the same time trivialises that notion.
- Re-examines some of the fundamental assumptions of educational assessment generally and psychometrics in particular. Indicates some of their most blatant self-contradictions and fudges.
- Reconceptualises the notion of invalidity, and positions the field of educational categorisation here, from the perspective of the examined, rather than with validity, which is an advocacy for the examiner.
- Applies some of this analysis to a study of competency standards in general, and in particular University grades, and national literacy testing as developed in the Australian context during the

1990s.

As can be seen, the initial research question has generated action as well as understanding, a tool to repair the damage resulting from the critique, and a way to reduce some of the violence it implies.

Relevant Literature

The relevant literature is extensive as well as intensive, as the Bibliography shows. The extensiveness was necessary, as many of the misconceptions and fudges and contradictions that characterise the field of educational assessment have been caused by a myopia regarding knowledge outside the arbitrary boundaries within which the field encloses itself.

Within the field of educational measurement the critical studies which most overlap mine are: in the United Kingdom, Hartog & Rhodes (1936), Cox (1965); in the United States, Hoffman (1964), Nairn (1980), Airasian (1979), and Glass (1978); in Australia, Rechter & Wilson (1968).

The Hartog & Rhodes study clearly showed the enormous instability of the measurement of standards in Public Examinations in England. The sneakiness of some of the research techniques in no way detracts from the dramatic incisiveness of the data. Cox did a similar job and ended up with a similar horror story on measurements of University grades. Hoffman directed his critical attention to the detail of multiple choice testing. Nairn's critique of the work of Educational Testing Service, and in particular the part it plays in College Entrance, is devastating in its implications. Airasian's book is a comprehensive critique of competency testing. Glass attacks the measurement of standards at its most vulnerable point; there are no standards, or at least none that psychometrics can produce. And Rechter & Wilson's study indicates the confusion about how to reduce error that accompanies public examining in Australia.

On the other hand, most of the literature on reliability and validity is pertinent to this study, because, when its discourse is repositioned from examiner to examined, it provides more than enough invalidity information to self destruct.

Most studies of error in the measurement of standards are however much more specific in their focus than is mine. Their minimal effect on practice has perhaps partially been due to the fact that their critiques were in terms of their own discipline of educational measurement; a discipline that owes its very existence to the claim to accurate judgments. In terms of general style and scope this study is perhaps closer to the work of Persig (1975; 1991), who delved, articulately if deviously, much more deeply into the notion of quality.

Within the field of power relations and the construction of the individual the studies most similar are those published in Foucault and Education (Ball,1990), in particular those that take off from Foucault's placement of the examination as a central apparatus of power/knowledge.

This study is significant in that it brings these two diverse fields of educational assessment, and the power relations that pervade education, into much closer contact, to expose their interrelations, and allow the critique to cross fertilise.

Importance of the study

The initial question addressed is how the whole matter of error in measurement of standards is obscured in most practical events involving assessment and measurement.

This is directly related to the centrality of the notion of the educational standard to the categorisation, production and control of the individual in society. For if the notion of the standard is crucial to the maintenance of power relations, and its empirical realisation is prone to enormous error, then the whole apparatus of power/knowledge that depends on it is in jeopardy.

I argue in Chapters 4 and 5 that the examination normalises and individualises, and is impotent without the notion of the measured standard, the sword that divides, the wedge that produces the gaps; and how important it is that these measures of standards be seen as accurate if current societal structures are to be maintained.

One view of immorality is that it is behaviour that destabilises a social system. So if playing the game is inevitable, is questioning the rules not so much dangerous as despicable, immoral to the point of being unthinkable? Is this the reason for the great silence about the enormous errors in any measure of standards? Does this account for the erasure from public consciousness and discourse of the obvious fact that educational standards as a thin accurate line have no empirical existence, and attempts to measure in relation to that line no instrumental reality?

In Chapters 6 to 17 thirteen sources of invalidity that contribute to the error and confusion of all categorisations of individual persons are detailed and elucidated, indicating how this silence in professional and public consciousness might be filled with a deafening noise.

In Chapters 18 and 19 of this study I apply some of the analytic tools developed to the contemporary scene in Australia, and demonstrate how the noise may be turned into a coherent critique of practice. In 1997 competency standards, as a form of assessment, have become, and are becoming, the major credentialing instrument for both educational and vocational courses and jobs. In addition, they are now the basis for job descriptions. In defining what training is required for a job, what prerequisites are required to attempt a job, what the job is, and how performance on the job is to be assessed, the cycle of fantasy created by this controlled semantic reductionism is complete; the material world of education and employment has become textualised in terms of competencies (Collins, 1993; Cairns, 1992). The fragility of this theorising is exposed when examined in terms of the reconstructed notion of invalidity developed in this study.

In Universities students are still categorised in terms of grades loosely defined. What do they mean? How error prone are they? And in the schools all Australian states have agreed to introduce tests of literacy. Certainly they will introduce tests. But what will they measure? And with what accuracy? Again the reconstructed notion of invalidity is used to critically evaluate such questions.

Methodology and the critique of practice

The study roves beyond the artificial constraints of psychometric theory and test practice; into ontology, epistemology and the metaphysics of quality; into the nature of instrumentation; into the relations between equity and assessment frames of reference; into the fundamental notion of comparability; into the detail of the relation between rank orders, standards and categorisations; and into the minefield of the psychometric fudge.

Is there method in this diverse madness? Where is the methodology that informs this wild profusion? The study aims to expose the madness that underlies much of the current method. So what is a methodology that undermines methodologies?

One such method is critical analysis, the analysis of the educational discourse that comprises the field of assessment. The polices and practices of educational assessment become fused in the discourse in which they are embedded (Ball, 1994).

> Discourses are about what can be said, and thought, but also about who can speak, when, where and with what authority. Discourses embody the meaning and use of propositions and words. Thus, certain possibilities for thought are constructed . . . We do not speak a discourse, it speaks us. We are the subjectivities, the voices, the knowledge, the power relations that a discourse constructs and allows (p22).

Analysis of such discourses may not be used to determine the truth. Yet such analyses may be very sensitive to the uncovering of untruths, by determining the extent to which they embody "incoherencies, distortions, structured omissions and negations which in turn expose the inability of the language of ideology to produce coherent meaning" (Codd, 1988, p245).

How would such untruths be established?

- First, by uncovering self contradictions, within the overt discourse, or between the unstated assumptions of the discourse and the facts that the discourse establishes.
- Second, by exposing false claims, claims that may be shown with empirical evidence constructed within its own frame of reference to be untrue.
- Third, by detailing some of the psychometric fudges on which many assessment claims depend to maintain their established meaning.
- Fourth, by indicating how repositioning the discourse may dramatically change its truth value.
- Fifth, by establishing four discrete epistemological frames of reference for assessment discourse as currently constructed, and

indicating the confusion when one frame is viewed from the perspectives of the others.

- Sixth, by noticing frame shifts within a particular discourse, with the resulting confusion of meaning.
- Seventh, by exposing the ontological slides and epistemological camouflages necessary to sustain many truth claims.

So in this study I will substantiate the contention that some of the explicit and implicit "truths" embedded in assessment practices are falsifiable; that empirical data constructed from their own assumptions denies the accuracy they assume; that this data is not only adequately detailed in the literature, but further, that the notion of error is the epistemological basis of much of that literature. All of which makes the public silence about the presence of error even more puzzling.

I shall show that the epistemological and ontological grounds for the whole field of assessment of individual persons are enormously shaky. I shall also explain how the literature about the very notion of validity is founded on a biased position, so that the sources of invalidity are much deeper and wider than is admitted in practice, even though clearly implied in theory and its attendant discourse.

I shall indicate the complexity of the notion of invalidity, with its practical face of error. Error includes all those differences in rank ordering and placement in different assessments at different times by different experts; all the confusions and varieties of meaning attached to the "construct" being assessed; and all those variabilities arising out of logical type errors, issues of context, faulty labelling, and problems associated with prediction. To further complicate the matter error has a different meaning depending on the assessment frame of reference. And I will show that estimates of the extent of the confusion along many of these dimensions may be easily estimated.

This is a critical study. Foucault (1988) says:

> There is always a little thought even in the most stupid institutions; there is always thought even in silent habits. Criticism is a matter of flashing out that thought and trying to change it: to show that things are not as self-evident as one believed, to see that what is accepted as self-evident will be no longer accepted as such. Practising criticism is a matter of making facile gestures difficult (p155).

Using Foucault's terminology, this is a critical study designed to make facile assessment gestures about standards difficult.

Methodology and inquiry systems

After a twenty three page discussion on data and analysis relevant to construct validation, which to Messick (1989) means all validation, he concludes

> . . . test validation in essence is scientific inquiry into score meaning - nothing more, but also nothing less. All of the

> existing techniques of scientific inquiry, as well as those newly emerging, are fair game for developing convergent and discriminant arguments to buttress the construct interpretation of test scores (p56).

I would broaden this to refer to any categorisation produced by transforming a continuity into a dichotomy. And for now I want to leave aside the obvious bias in the word "buttress," and focus here on inquiry systems themselves. For Messick (1989), conservative as he is, accepts that

> because observations and meanings are differentially theory-laden and theories are differentially value-laden, appeals to multiple perspectives on meaning and values are needed to illuminate latent assumptions and action implications in the measurement of constructs (p32).

Churchman (1971), elucidates five such scientific inquiry systems of differential values and epistemology, roughly related to philosophies espoused by Liebniz, Lock, Kant, Hegel and Singer. Mitroff (1973) has developed and summarised Churchman's systems. Very briefly, the Liebnizian inquiry mode begins with undefined ideas and rules of operation, ending with models that count as explanations. The Lockean mode begins with undefined experiential elements, and uses consensual agreement to establish facts. The Kantian system shows the interdependence of the Liebnizian and Lockean modes, and uses somewhat complementary Liebnizian models to interrogate the same Lockian data bank, to ultimately arrive at the best model. The Hegelian mode uses antithetical models to explain the same data, leaving it for the decision maker to create the most appropriate synthesis for a particular purpose. In this mode values of enquirer and decision maker become exposed. Finally, the inquiry system of Singer (1959), is one of multiple epistemological observation, where each inquiring system is observed from the assumptions of the others, and each methodology is processed by those of the others. Churchman (1971) paraphrases Singer clearly and cleanly: "the reality of an observing mind depends on it being observed, just as the reality of any aspect of the world depends upon observation" (p146).

How do these inquiry systems link to the seven ways of demonstrating untruths, or nonsense, detailed in the previous section? It is the Singerian inquiry mode that best characterises this study as a whole. Although particular modes have been utilised for particular critical purposes, this is in itself justified by the Singerian inquiry mode.

So whilst the first three methods listed are clearly in the Liebnizian and Lockean modes, the other four involve the explication of shifting sets of assumptions, and belong to the Singerian mode. In particular the examination of compatibilities between the four frames of reference for assessment on the one hand, and equity definitions, power relations, instrumentation requirements, and notions of comparabiltiy and quality on the other, demonstrate clearly that to the

Singerian enquirer, "information is no longer merely scientific or technical, but also ethical as well" (Mitroff, 1973, p125).

The "conversation pieces" and "stories" used to demonstrate the absurdity of some assessment claims belong to the Hegelian mode. Churchman (1971) explains:

> The Hegelian inquirer is a storyteller, and Hegel's thesis is that the best inquiry is the inquiry that produces stories. The underlying life of a story is its drama, not its "accuracy". Drama has the logical characteristics of a flow of events in which each subsequent event partially contradicts what went before; there is nothing duller than a thoroughly consistent story. Drama is the interplay of the tragic and the comic; its blood is conviction, and its blood pressure is antagonism. It prohibits sterile classification. It is above all implicit; it uses the explicit only to emphasise the implicit (p 178).

Strategy of deterrence

The general strategy used to make the case for the invalidity of most current assessment practice is borrowed from military policies of nuclear deterrence. It is a strategy of overkill. Of the thirteen sources of invalidity developed in this study, any one would, if fully applied to current assessment practices, take them out, neutralise them, render them inoperable. To nullify this attack on validity of tests, examinations and categorisations generally, it is necessary to destroy not one missile, but all of them.

Methodology and structure of the study

The study has been presented in seven parts: Positioning, Context, Tools of Analysis, Error Analysed, Synthesis, Application, and a Concluding Statement.

Part 1 - Positioning : All descriptions of events, all writing, is positioned; makes certain assumptions, is viewed from a particular perspective. Part one positions the study in terms of focus and method, and the writer in terms of experience and philosophy.

In this opening chapter I position the work in terms of its general content and methodology, and show how it all fits together. So Chapter 1 briefly summarises what the study is about, what literature is most similar in both content and style, what is the importance of the study and its possible impact, and in this section how it is structured.

In Chapter 2 I show how the study is positioned in terms of some of the learnings accrued from the professional and life experiences of the author.

In Chapter 3 I indicate how the study is positioned in terms of philosophy and value, and how that relates to some contemporary literature.

Part 2 - Context: Assessment involves events that occur in, and are given meanings in, a social context. In Part 2 I elucidate some aspects of that context.

In Chapter 4 I focus on the way power relations both violate and produce those who act out their lives within their influence. In particular the centrality of the examination is exposed in the production of the modern individual, defined as an object positioned, classified and articulated along a limited set of linear dimensions.

In Chapter 5 the argument in Chapter 4 is applied and developed in terms of educational assessment. In particular I examine the crucial part that the standard plays in the whole mechanism of defining cut-offs for abnormality and non-acceptance, and how important it is that these standards be seen as accurate if current societal structures are to be maintained.

In Chapter 6 I focus on the cultural meanings that attach themselves to the notion of the standard, and assign the idea of the human standard to the mythological sphere, a place apart from critical thought. I examine the emotional intensity of discourse about the standard, its significance as an article of faith, and how this is related to the maintenance of control and good order.

Part 3 - Tools of analysis: In Part 3 some tools for looking at specific assessment events are developed. In Chapters 7 to 12 I examine four different epistemological frames of reference for assessment, and relate these to notions of equity, to hierarchical structures, instrumentation, comparability, rank orders and standards, logical types, and quality. These chapters introduce some independent, fundamental, and rarely discussed aspects of underlying assumptions involved in events culminating in the assessment of students. Inadequacies in any one of these aspects would, in a rational world, be enough to destroy the credibility of most student assessments. I will contend that all practical assessments of people contain major inadequacies in most of them.

In Chapter 7 four different frames of reference are defined; four different and largely incompatible sets of assumptions that underlie educational assessment processes as currently practised: First is the Judges frame, recognised by its assumption of absolute truth, its hierarchical incorporation of infallibility; second is the General frame, embedded in the notion of error, and dedicated to the pursuit of the true score; third is the Specific frame, which assumes that all educational outcomes can be described in terms of specific overt behaviours with identifiable conditions of adequacy; fourth is the Responsive frame, in which the essential subjectivity of all assessment processes is recognised, as is their relatedness to context.

Because of their contradictory assumptions, slides between frames result in confusion and compound invalidity.

Chapter 8 shows how certain assessment frames are inherently contradictory to certain definitions of equity, themselves contradictory to each other and to the power structures in which they are enmeshed. As such, those assessment frames and notions of equity that contradict the enveloping hierarchical structure will be seen, accurately and probably unconsciously, as potentially destabilising, and will consequently be ignored, nullified, or corrupted into acceptability.

Chapter 9 looks at Instrumentation. In this chapter we look at the conditions and invariances required in events involving measuring instruments if such events are to have credibility; in particular the notion of a Standard that theoretically defines the scale, and its confusion with a standard of acceptability, which is to be measured by the instrument, and which requires a scale in order to be located.

The various assessment modes are analysed in terms of their instrumental error. On these grounds alone all are found to be invalid.

Chapter 10 takes up the issue of comparability. What can be compared? Fundamental distinctions between more and less, better and worse are examined , their relations with uni and multi dimensionality shown, and the implications for rank ordering of students in tests and examinations unearthed. This leads to further examination of the differential privileging of sub groups and individuals when marks are added. The essential meaninglessness of such additions becomes apparent.

In Chapter 11 the relationship between rank order and standard is teased out in more detail: In particular the meanings given to the standard in the Judge and General frames of reference; how logical confusions proliferate when discourse jumps from one frame to the other; and how all categorisations involve standards and rank ordering, even though many advocates of "qualitative" assessment methods may want to deny this.

Chapter 12 leads from the implications of the Theory of Logical Types for assessment practices to an examination of the distinction between standard and quality. When the standard is seen, realistically, as unable to perform its function, quality is the notion with sufficient mythical, ideological, and intellectual status to replace it. This would produce a very different learning milieu.

Part 4 - Error analysed: In Part 4 the tools developed in Part 3 are used to discriminate particular sources of confusion and error within assessment events designed to categorise students.

In Chapter 13 the meaning of error in each frame of reference for interpreting assessments is considered. As the meaning of error changes with assessment mode, so do the methods designed to reduce such error. Procedures to reduce error in one frame are seen to increase it in another. From a perspective of oversight of the whole assessment field, this is another source of confusion and invalidity, particularly as it is rare for any practical assessment event to remain consistently within one frame of reference.

Chapter 14 addresses the question: What does a test measure? In terms of social consequences the answer is clear. It measures what the person with the power to pay for the test says it measures. And the person who sets the test will name the test what the person who pays for the test wants the test to be named. The person who does the test has already accepted the name of the test and the measure that the test makes by the very act of doing the test. So the mark becomes part of that person's story and with sufficient repetitions becomes true.

My own conclusion is that tests have so many independent sources of invalidity that they do not measure anything in particular, nor do they place people in any particular order of anything. But they do place them in an order, along a single line of "merit," and that is all they are required to do.

Chapter 15 shows some of the ways in which psychometricians fudge; by reducing criteria to those that can be tested; by prejudging validity by prior labelling; by appropriating definitions to statistical models; and by hiding error in individual marks and grades by displaced statistical data, and implying that estimates are true scores. A number of specific examples of fudging are detailed.

In Chapter 16 some of the more recent work on validity is discussed, and its positioning as advocacy demonstrated. I conclude that in practice the very existence of validity is established, validity is indeed made manifest, through the denseness of the arguments about invalidity criteria used to refute such existence, together with the reassurance that the battle continues, and some gains have been made.

Reliability is also discussed as a problematic, rather than as an obvious prerequisite to validity. I conclude that most of the mechanisms designed to increase reliability necessarily decrease validity.

Part 5 - Synthesis: In Chapter 17 the notion of invalidity is reconceptualised, having both discursive and measurable components. Thirteen (overlapping) sources of error are examined, all contributing to the essential invalidity of categorisations of persons.

Part 6 - Application: In Chapter 18 I apply the philosophical and conceptual positioning, tools of analysis, and the reconceptualised sources of error developed in this thesis to the competency based assessment policies and practices of Australia in the 1990s. I show how the notion of competency standards is overtly central to the whole competency movement, the introduction of which is shown to be overtly politically motivated. Thus the crucial links between political power and educational standards that are argued for in Chapters 3 and 4 become transparent. I then go on to examine the invalidity of competency standards in the light of the thirteen sources of error specified in the previous chapter.

Chapter 19 presents two specific applications of invalidity sources; the first relates to national literacy testing, and the second to University grades.

Impact

Assessment practice is permeated with mythology and ideology; with confusions and contradictions; with epistemological and ontological slides; with misrepresentations of frames of reference for different assessment modes; with logical type errors and psychometric fudging, in which the constructs that determine error--labelling, construction, stability, generality, prediction--are either ignored or severely constrained in the determination and communication of error, in those rare cases where personal error and likely miscategorisation is publicly admitted.

I have no expectations for this study, but some hopes. A whistle blowing study is like a joke--its impact is a function of timing. And the best timing can only be determined in retrospect. My hope is that it will lead to a reduction of the violence that is attributable to the suppression of error in the categorisation of people.

Return to Table of Contents

Chapter 2: Positioning the writer: experience

Introduction

As I take the epistemological position that all knowledge is based on experience, value and reflection, and all experience is influenced by prior knowledge, it seems important to indicate some of those life experiences that led me to the particular ontological and epistemological positions that inform this study. To do otherwise is to infer either their universal superiority, or their complete arbitrariness.

In this brief autobiographical note I outline some of those significant life experiences and concomitant learnings as they impinge on this study. This is neither arrogance nor self-indulgence (Mykhalovskiy, 1996). For if thirty years working in the field of educational research and assessment is not relevant to this project, then either the work, or the project, or both, must surely be trivial.

Education

This study has had a long gestation. Forty nine years ago I sat for my matriculation examination in English. I had a choice of four essays, and chose one called "Examinations." I rubbished them, unwisely it seems. I got a B grade which compared unfavourably with the second highest mark for English at my prestigious public school. That I'm still at it today indicates that non-conformity is not necessarily related to inconsistency or nonperserverence. What I learnt from this experience is that meaning and judgment are affected by context, and that appropriateness is one criterIon for the recognition of quality.

Two years of study in the University Engineering faculty convinced me that I did not want to be an engineer, and left me with one invaluable legacy; on every engineering drawing the measurement of each dimension, and the limits of accuracy within which the product must be fabricated, are indicated. In practice, because error was inevitable, the statement of acceptable error was as important as the magnitude of the dimension. Keeping within acceptable error was a major determinant of quality of product. This practice of indicating errors in measurement continued for calculations in Physics, the subject of one of my majors when I transferred to the Science faculty.

I decided to become a teacher. Moving to Education was a culture shock. I could only write scientific prose - sparse and unadorned, tight and dry, logical and on the surface devoid of any emotional involvement. So writing two thousand word essays was a problem; I generally said all I had to say in two hundred, and regarded the rest as superfluous padding. I could state my case, but had lost my personal voice.

What I learnt about assessment was at the level of "helpful hints to beginning teachers." The massive literature on educational assessment and evaluation was then, as it is now for most teachers, unknown to me. I was trained for survival, not for problematising tradition. I learnt what was implied. The game of testing had produced me, so it couldn't be all that bad.

Teaching

I taught in high schools and tested students more or less the way I'd been tested. Maybe a few less essays and considerably more short answer questions. The process was simple. I sat down, wrote some questions to comprise an examination paper, the students did it, I marked it, added up the marks, and then gave them a percentage or converted it to a grade. How was it done? Easy! Was it a problem? No! How accurate was it? Nobody, including me, ever asked!

After three years I joined the Royal Australian Air Force as an Education officer, teaching some basic physics to photographers, some nuclear physics to air crew, and some instructional technique to officers. Because I was teaching it, I learnt the technology of lecturing. It was assumed I could accurately assess all this. I averaged about six lectures a week, so they were very well prepared. With so much time, I diverted myself by writing pantomimes and musicals. I was beginning to find my voice.

Two years of work at the RAAF School of Technical Training had me writing syllabuses as well as teaching basic maths and physics. I talked to electrical fitters who had come back for training after two years in the field as electrical mechanics. None of them had used any of the eighty odd hours of mathematics in the Mechanics course. I suggested to the administration that they save time and money by leaving out the mathematics. It was explained to me that its relevance to work was irrelevant. It was necessary for the high level of trade classification. I was beginning to understand the economic and political character of credentialing.

Assessing

My last year in the RAAF was spent in the trade testing section. Fifty item, two hour, multiple choice tests were used to credential students who had spent from three to twelve months in training programs, with hundreds of hours of practical and theoretical assessment as part of the course. My attempts to point out the absurdity of this were usually met with the response that it didn't matter, because they just kept on doing the trade tests till they passed. I was becoming aware that in the world of work, as well as in the world of education, ritual was more important than rationality.

Teaching again

Observing that the influence Education Officers had on training seemed to diminish as they were promoted, I went back to teaching in a private coeducational high school. I found that what had taken twenty hours to teach to highly motivated technicians took five times as long to teach to supposedly more intelligent high school students. In my second year I told the matriculation physics class I did not intend to teach them. Rather I would try to create an environment in which they could learn. I would assume they could read the syllabus and the text book. They worked individually or in groups, developed their own notes, devised their own experiments. They completed the course by the end of June, after which I agreed to give some consolidating lectures, and class time was spent doing past examination papers and improving answers. That, after all, was the task on which they would be judged. Their results in the external examination were extremely high. I had learnt to separate the ritual of teaching from the facts of learning.

Next year I tried the same process. The students refused to cooperate. They collected notes from other schools. They insisted I teach them. After a month I had little choice. We went back to "normal" teaching methods. They got "normal" results at the end of the year. I learnt that dependency has as much attraction as autonomy, for the price of autonomy is personal responsibility.

Two other events were significant over this period. The first was a question asked by Michael, a student; What exactly is an electron? I had no idea. The question had never occurred to me. I'll let you know, I blustered. A month and many hours of reading later, I responded. Do you remember, Michael, you asked me what an electron is? No, he answered. I'll tell you anyway, I said, unperturbed. I wrote "Properties of an electron" on the blackboard, and under that heading listed some of them. The class looked on in silence. I looked at Michael. Yeah, he said, those are its properties, but what exactly is it? Ah, I said, now that's a question you'll have to ask the Rabbi. I had started to grapple with ontology. I was thirty years old.

Writing

The second involved the writing of A programmed course in Physics (Wilson, 1966). This was a linear program covering year 11 and 12 Physics. In reviewing what I had written I was dissatisfied with the presentation of force field theory. Finally I wrote this part as a dialogue between a physicist and a student. The result was much more satisfying in that the nature of a field in physics could be discussed as a problematic, rather then presented as a scientific conclusion. My first excursion into epistemology required discourse rather than didactic prose to communicate its meaning.

Assessing again

Because of my experience with multiple choice tests in the RAAF, I had been working with Australian Council for Educational Research on the construction of multiple choice physics tests. When a full-time position came up I applied for it. For the next six years I was to work as a test constructor. I learnt a lot about the nature and mechanics and rituals of testing, about the truisms and tricks of the trade. For example, that only "items" between thirty and seventy percent difficulty were chosen because others did not contribute economically to the separation of students; that seemingly almost identical questions often had very different difficulty levels; and it was almost impossible to tell, without prior testing, how difficult a test item was.

Central to the theme of this study, I also learnt, at the level of practice and praxis, the great secret about error, about the fallibility of the human judge, about the vagueness and arbitrariness of the standard. Not in that language, of course. Psychometrics provides a more prophylactic discourse about marker reliability and predictive validity and generalizability. Even so, it was impossible to miss the point. Or was it? I did a course in educational measurement at a local university to sharpen up my theoretical skills. We learnt the statistical theory and all the little techniques for reducing error, like short answer questions and multiple marking. And at the end of the course--a three hour essay type examination marked by the lecturer and then given a grade. And nobody said a word! Even more amazing, when I raised the matter with a few of the other students, they seemed unaware of the contradiction. I was learning that tertiary studies do not necessarily invoke reflective critical thinking.

There were two other outcomes of this experience of constructing test items that were important. The first related to the discourse, the arguments about the best answer that characterised the panel meetings. The second related to the values and effects of this particular testing program, and how to deal with that (Wilson, 1970).

As we got better at writing "distractors" for multiple choice questions, we found advocates among the "expert" panel for some of the distractors as the best answer, rather than the one chosen by the test writer. Of more potential educational significance was the argumentation itself, and its effect on our ability to think sharply and clearly within the fields being discussed. Tests themselves can never produce improvement in individual performance; but our experience suggested that argumentative discourse about test items could. A serendipidous piece of research at one school confirmed this. One hundred students thus engaged for about twenty hours raised test scores on each of three multiple choice papers by half a standard deviation, despite the ACER publications that claimed these tests could not be "taught" (Wilson, 1969).

The second experience related to educational values, and our attempts as "examiners" to grapple with this. None of the full-time test constructors approved of the Commonwealth Secondary Scholarship tests as an educational intervention. They were a politically inspired election gimmick. We were aware that they would have an influence on what schools taught, and possibly how they taught, even though they were supposed to be "curriculum free" as well as value free. As a result we took "educational value" as a major criteria for test validity, at least at the level of our own personal discourse. The material we chose for tests must face the question "would education be improved if teachers did try to prepare students for this sort of exercise, for answering these sorts of questions on these sorts of information or issues, for engaging in this sort of thinking and problem solving?" I was learning that no test was value free, and that these tests were certainly informed by a (possibly idiosyncratic) view of educational relevance.

Groups

During these years I also had my first experience in unstructured groups, and experienced at first hand the power of such group interactions to produce major changes in social behaviour in the participants; within the microcosmic society of such groups, as they developed, there was opportunity to take risks, revisit social experience, and re-construct social meanings. I learnt how powerful such groups could be in raising awareness, loosening counterproductive behaviours, and reframing experiential meanings (Slater, 1966).

Research

When at age forty I was appointed to head the newly established Research and Planning Branch in the SA Education Department, a position I held (with planning dropped half way through), for the next thirteen years, my major claim to expertise was in the area of testing and assessment. The Directors never allowed this to influence their decisions about committee membership, and during my sojourn with them I was never appointed by them to any departmental committee concerned with assessment. Nor, for that matter, am I aware of any decision made by the Department that was informed by research that the Branch carried out. When research knowledge was consistent with Departmental policy assertions it was utilised; when it didn't or wouldn't serve those interests it was ignored. I was learning that research knowledge was an instrument of power, a weapon for rationalising decisions, rather than a springboard for rational decision making (Cohen & Grant, 1975).

It was partly this insight, as well as a belief that my clients were students and teachers rather than administrators, that determined that most of my own research would be concerned with classroom practice. I also noticed that most educational research dealt with special groups and special problems, leaving the "normal" educational assumptions and practices unsullied by any critical research probes. So I directed most of my action research to the "average" classroom; that is, I sought out the commonalities of educational experience rather than the differences.

In the first few years I spent considerable time with teachers looking at improving assessment practices in schools. One thing in particular became apparent during these discussions--that most of what I had learnt as a professional test constructor was irrelevant to the assessment issues that concerned teachers in classrooms; these were not the sort of descriptions that helped children learn better, or helped teachers teach better. When I wrote Assessment in the primary school in 1972 the then Director of Primary Education wrote a foreward in which the final paragraph stated "some people would question his suggested limitation on testing. Whatever one's views, teachers will find the report thought provoking and valuable". In other words, I disagree with him, but respect his different viewpoint. As Directors became more managers and less educators in the 1980s, this sort of clarity and openness, this up front honesty, was to become increasingly rare.

Politics

In 1974 a thirteen year old schoolgirl was suspended from her high school and refused to accept the suspension on the grounds that it was unfair. She returned to the school and was subsequently removed forcibly by police. The incident resulted in a Royal Commission, and the Royal Commissioner found that the girl and her parents were a "trinity of trouble makers". (Royal Commission, 1974). It was never suggested that the setting up of the Commission had anything to do with the fact that the girl's father was an endorsed labour candidate and a personal friend of the Minister of Education, and that the Principal of the school was the brother of the shadow Minister of Education. Nor was it ever suggested that the united front of the Education Department officers and secondary principals had anything to do with the highly conflictual situation then existing between the Minister and the high school principals.

I thought that most of the overt conflict at the school was due to communication problems between the girl and certain members of staff, and certainly not due to the severity of the crime, which was trivial. In such cases it seemed to me to be the job of the professional staff, not the student, to resolve the conflict. So I gave evidence on behalf of the student. I was the only member of the Department to do so. What I learnt from this episode was that the structural violence embedded in institutions is evidenced not by the severity of the punishment when rules are breached, but by the severity of the punishment when the sanction, whatever it is, is not accepted. I could see that accepting any sanction reinstates the power structure; in fact, breaking the rule enables such re-establishment to become visible, enhancing the power relations. But not accepting the sanction is extraordinarily threatening because it destabilises the power structure, challenging its very existence. It also became clear to me that none of the Departmental officers, or the Royal Commissioner, could see this.

Social development research

As the development of social skills was a major objective in the stated curriculum of almost all school subjects, I initiated a major project on social development. It lasted four years, attracted two major grants, and at one stage involved six full time and six part time researchers (The Social Development Group, 1979). As a starter to this I took six months long service leave and a round the world trip. I spent some time visiting people and relevant projects in the United States, Canada, and England. I talked to teachers at primary, secondary, and tertiary levels about the social development of their students, and how they were able to facilitate that development. They all described the social development of their students during a year, whether six or twenty six years old, in the same terms; tentative, inarticulate, immature to confident, articulate, sensitive. It was obvious that what they were talking about had little to do with developmental skills.

My experience in unstructured groups suggested to me that it had everything to do with developing groups, with the way that power, affect and trust relations change if they are allowed to. I had already spent six months reading the literature on social skill development. It was often interesting, but utterly uninformative in regard to classroom practice. And we had asked teachers to describe mature social skills; they responded with good descriptions of conforming behaviour. I could see that shifting the focus to the social group, to the context of social action, produced an array of possible teacher interventions, informed by group development theory. We started with a project about developing social skills. We ended with a project on developing the classroom group; for only in a developed group would the demonstration of mature social skills be appropriate.

Rebelliousness

One incident that occurred on this journey deserves a mention, as it relates to the question of what constitutes experience. In London I went into a coma for two weeks, during which time I convulsed and hallucinated and was fed by a drip and lost 12 kilograms in weight. I was diagnosed as having viral encephalitis.

My hallucinations had a clear story line. They all involved adventures with semi humanoid monsters who were trying to kill me. The final scene had me lying on an operating table with ten humanoid gun barrels at my head. The odds were stacked against me, and death was immanent. I had time only for one statement. "You will only kill me," I said, "to prove that I cannot control you. Yet if you kill me for that, then I have completely determined your actions." They left, I came out of coma, and requested some food. With some trauma, I had learnt that the rebel is as tied to the system as the conformist. If I wanted to change the system, I would have to take a different stance; one of autonomous action, rather than rebellious reaction. I would need to tap the ambivalence of those in power, not their antagonism.

Back in Adelaide, the social development project got under way. I read the literature on (small) group development theory, and realised that most of the models could be reframed in terms of distributions of power and affect relations; and because of my physics background, I conceptualised these in terms of fields; properties of the space between rather than of the agents mediated by the fields. My personal ontology was developing, and ten years later more complex notions of power relations (eg Foucault) would find nourishment in my conceptual space.

Politics again

Part of the condition of the research grant was that separate reports be written for the major participants in the study; researchers, administrators and curriculum writers, teachers, students. I wrote the booklet for students. It was entitled How to make your classroom a better place to live in (The Social Development Group, 1980). It described the four stages of development of the classroom group, how students might experience these stages, and how they might respond to that experience. Four different responses to each situation were constructed, and were overtly categorised as positive and negative; the negative responses, with which students would identify and be familiar, were likely to be not constructive in moving the group onward; the other two responses, one involving individual action and one group action, were ones which might help the group develop. The booklet was designed for classroom discussion.

Before the book was distributed a question was asked in the South Australian parliament about the book. Was it not encouraging students to respond negatively? The Director General responded by ordering that the book be shredded. Flattered if furious with this treatment, I pointed out the conditions of the grant, and requested specific information about exactly what was objectionable in the book, so that it could be amended and reprinted. After some months the answer came back; two words, "fascist" and "fairy," had to be removed; the positive responses must come first; and there must be an overt statement that the positive responses were "better". In addition, only teachers involved in developing their class groups could distribute this book to their students.

I interpreted this to mean that there was nothing specifically at fault with the book. It was the ideology of the book, with its implicit aim of empowering students, that had caused the over-reaction. Yet the rhetoric about schools applauded the empowerment (autonomy) of students. Unwilling to confront the contradiction, the Department had to settle for limitation rather than complete suppression. For of course developing the classroom group meant that the power relations between teachers and students changed. If this happened in enough classrooms not only classroom structures, but school structures, would have to change. The implications of the research were radical rather than progressive.

Inservice training was essential if the findings of the research were to be propagated, if practice were to follow theory. So four researchers, now highly skilled in working with teachers, were retained for a year to produce inservice materials and work in schools with teachers. A year later, despite protestations, all had been returned to classrooms. An invaluable human resource for the dissemination of ways of developing the classroom group was annihilated. Fifteen years later teachers still struggle with rebellious classrooms and search for answers in individual psychology, curriculum statements still highlight the development of social skills rather than the social context for mature social behaviour, and teachers still say "groups don't work" because they don't understand group development theory. In 1980, I was beginning to learn what I knew by 1990; that nothing really changes unless the power structure changes, and hierarchical power structures are immensely stable and resistant to change (Wilson, 1991).

Consciousness

One further event in 1979 is pertinent to this story. At Findhorn, an intentional community in Scotland, I experienced some shifts in consciousness (without drugs or intention, with detachment and interest), that seemed very similar to those experiences described by mystics, and generally described under the rubric of the perennial philosophy. (Bucke, 1901; Huxley, 1946; Wilbur, 1977,1982,1991; Wilson, 1992). These experiences, and subsequent ones, make it impossible for me to take Freud's easy way out (Freud, 1963), and discount such events because I have not experienced them. Such experiences have been immensely significant in the history of the past three thousand years, for they have provided the bases for the world's great religions. The mythologies and structures that are the social manifestations of these initiating mystical events have taken very different cultural forms, but all have retained, within their core practices, considerable congruency with their source as a particular state of consciousness. This is important because it points to one exit from the maze of confusion created by the acceptance of the relativity and cultural determination of all human values (Wilbur, 1995).

Peace and violence

By 1982, Ronald Reagan's unique combination of monstrous stupidity and apocalyptic hardware had stirred the coals of fear still glimmering under the weight of twenty years of psychic numbing and denial, of human refusal to seriously consider the high probability of a nuclear holocaust that could destroy all life on the planet. Everywhere the peace movement flourished. Learned journals of all sorts from medicine to engineering, from physics to art, began to feature articles about nuclear war and its effects. Most unlikely bedfellows, Marxists and churchmen, pacifists and retired admirals, feminists and builders labourers, would all shout out their protests.

Where were the children in all this? I decided to find out. There was some American data from surveys. I decided to tap a richer source; children's fantasies of the future. The data was devastating (Wilson 1985). For many it was a post-nuclear war world, barren landscapes and destruction everywhere. For nearly all it was dehumanised, people existing either as passive recipients of technology, at the best comfortably mindless in a plastic world, at the worst slaves of the machines or robots that grind mercilessly along their efficient and pre-programmed paths. An unstoppable high-tech, high-destruct world.

Like many who start with a naive view of peace as the absence of war, my reading and reflection soon led to more sophisticated understandings; towards peace as the absence of fear at a psychological level, and as incompatible with injustice and repression at the social level. And I began to understand how injustice was often not so much a matter of human intention, as a product of historical man-made structures, continually reproduced through the human facility of role-taking, and the moralities and ideologies that are able to transform efficient violations into noble virtues. At fifty I was beginning to articulate a world-view.

During the international year of peace, schools were all expected to get involved. Believing that in dealing with violence we should begin in our own back yards, I prepared a kit for schools entitled Programs to reduce violence in schools (1986). It included ideas for involving students, teachers and parents, for collecting information, and for taking action at a school level. It also included a paper on understanding violence, in which I tried to make overt the links between violence, school structures, social control, and justice. Complete with words of encouragement from the Director General of Education, the kit went off to one hundred high schools in South Australia. One school got the project off the ground and collected data from students and staff. Then they stopped. During the year, many schools planted trees for peace. I was developing a feel for the absurd.

Writing again

Two years before, buttressed by a report by the head of another educational research organisation, the Department disbanded ours. I was sent out to graze in the country at Murray Bridge for two years as an Assistant Director Curriculum, where I managed to get two of the social development advisers back into business, before I retired gracefully. There was nothing further I could do within the system. I was ready to write, and had two young daughters at home that I wanted to spend more time with. I was learning the difference between jousting with windmills and hitting my head against a brick wall; one is a noble quest, the other just plain masochism.

The writing and the daughters got together into a book called With the best of intentions (Wilson, 1991). The book deals with the structural violence embedded in the hallowed institutions of family and school. I had decided to self-publish the book before I began, and as a result was able to give clear reign to my personal voice(s) and style. The book is egalitarian in that it treats children as fully human persons; it is iconoclastic in that it challenges many of the sacred myths and structures of child-rearing; it is written with passion and humour. It is informed by empirical data and overt in its philosophical world-view. The arguments are dense, but the presentation is, I hope, sufficiently varied and light to make its message accessible. With modifications that are essential to the context, I hoped to use a similar approach in this thesis.

The current study

A large number of significant learnings have emerged for me from the current study. I want to refer to the two that I have found the most significant. The first relates to my extensive reading of Michael Foucault, the second to my grapplings with ontology.

There were two major insights from Foucault; the first was his analysis of how culture produces and expresses rather than reduces and represses; that if the person is one dimensional, this is not because society has taken away the other dimensions, but that society, through its relations with the person, has produced a one dimensional person.

The second insight was the centrality given to the examination, in all its forms, to the construction of the individual in the modern world. It was from this springboard that I could leap to observe the standard as the bullet in the examination gun.

An equally important learning from Foucault relates not to insight, but to style; not to his immense data base and sometimes lugubrious argumentation, but to the soaring rhetorical passion that marks his insightful conclusions; his demonstration that "scientific" writing does not need to be dull and portentous, but can legitimately use the full creative resources of the language, helped me to feel much more comfortable in using my own voice for this work.

My own philosophical gropings into what is knowable, what is describable, led to some surprising conclusions. Such delving was necessary, because any assessment is a description. In practice it is a description of a performance of some kind in context, even if in theory it purports to be a description of some attribute or quality of a person; this I had known for a long time. To move from here to the insight that all knowledge is a description of events involving a relationship between at least two elements, and thus to appreciate the slide made when the description is pinned to one particular element, represented a major reframing of much of my earlier thinking.

Summing up

There are at least five levels in all this: The events that I was a part of; the manifest behaviour that constituted my part of those events; my particular recall of that experience; the meanings I verbally constructed from that recalled experience; and the meanings and reactions that you, the reader, construct from all that.

Truth is not an issue here. Awareness and truthfulness are. I can only assert my truthful intentions. Regardless, the reader will make his or her own judgment about the value of the position from which they interpret me as coming.

Return to Table of Contents

Chapter 3: Positioning the writer: philosophy and value

Preview

In this chapter I spell out in more detail the philosophical stance that I take in this study, so that my assumptions about social life and social relations are up-front.

Whilst these assumptions are consistent with the learnings of the autobiographical sketch give in the last chapter, I have not felt it necessary, or advisable, to enter into any sort of justifying dialogue regarding my position. This is not a philosophical study, and I have always regarded justification as a loser's game.

So I have presented my philosophical position as a set of assertions with an internally consistent logic; I have briefly described the epistemological, ontological, and axionomic assumptions that have informed this study, and described how that position fits into current post-positivist, interpretivist, and post-modern paradigms.

The chapter ends with a brief outline of the assessment process constructed from my particular position.

Philosophical assumptions : What is knowledge? What is truth?

I will call an event any interaction where a change or a difference is observed or otherwise sensed (Bateson, 1979). Interactions involve some relation between elements of the event. Differences involve some relation between the elements, or the states of an element over time, that constitute the difference. So all events involve some relation between elements. And because all events involve a perception, so all events involve a perceiver. The perceiver may be automated as an instrument that senses the difference or reacts to or records the change. As Maturana (1987) expresses it, "Everything is said by an observer" (p65).

Any experience is experience (action, feeling, perception) of an event, either directly, or as recalled or as transformed in memory or action. So all experience involves relations. As all knowledge must finally depend on experience, all knowledge involves knowledge of relations; so all knowledge is constructed out of relational events.

To experience an event does not necessitate giving a meaning to that event, but does require a state of awareness or consciousness, from which the event is viewed. For example, an experience may be represented by a pattern or abstract painting which embodies relations without embodying meaning. Giving a meaning to an event requires some theoretical underpinning, some ideas or ideals; some knowledge of relations derived from other events, or possibly, if mathematical relations are construed to constitute meaning, derived from acts of imagination that transcend (are transformations of) known relations. Mathematics can be regarded as a special case of patterning, and whether mathematical propositions or systems have meaning in themselves is moot. I don't think they do. Some post-structuralists want to deny experience that excludes meaning and thus language. My experience denies their denial. Their assumptions refute my denial. Stalemate. But then, I'm writing this thesis.

I use the term meaning to involve more than prediction, which mathematics can sometimes help to accomplish. Meaning involves some reason, some purpose, some intention, some value. Thus meaning is inevitably embedded in language, itself embedded in human discourse. Unless we take a mystical view and define the meaning as the experience itself, or rather as a particular encompassing experience, in which case discourse stops and the world in its oneness pulsates. In this thesis I shall hold to the more mundane view. To do otherwise is not to proceed.

In this epistemology, experience precedes pattern, and pattern precedes meaning. "Whether we are talking about unicorns, quarks, infinity, or apples, our cognitive life depends on experience" (Eisner, 1990, p31). Meaning will then usually in its turn, but not necessarily, pre-empt and distort experience, which will then in its turn influence events. Buddhist meditation is designed to limit this distortion; which brings its participants on this issue close to post-positivists like Phillips (1990), who seem ultimately to define objectivity as the reduction of bias of various sorts.

Meaning is socially constructed because language is socially constructed. What passes for knowledge in common language is a social concurrence in a particular culture about acceptable meanings embedded in discourse. On the other hand, experience is constructed out of relational events not necessarily linked to any particular culture, and the construction of patterns or relations in response to that experience may also sidestep, or transcend, social patterning or common meanings. In other words, I hold the view that creation is immanent in all events, and in all perception of events, and change is more than the imposition of some random variation. Usually, however, we may assume that patterns are also culturally influenced.

Data is a particular form of knowledge constructed by particular people for particular purposes. Such purposes always involve the construction or isolation of events in which the observer is directly, or indirectly through associated theory, involved; for example, measuring devices involve the observer at one step removed. Thus all data, being knowledge, is constructed from events, constructed and/or observed for particular purposes. All data, to be used, must have either a predictable pattern, or a meaning, or both. So if data is to be useful, it must have links to other relational events, or have links to (uneventful) abstract relations.

It follows that, in this world, there are as many potential truths about an event as there are experiences of the event. To the extent that all experiences of the event are the same then there is a case for "the" truth. But how would this be known? Any attempt to know this would involve the sharing of meanings, which are certainly socially constructed and can be as varied as the cultures and relations and metaphors that are used to make sense of them and communicate them. So agreement about one meaning, one truth, represents conformity about social construction as much as it does concomitance of experience.

Ironically, in a social context the idea of multiple truths is unificatory, whilst the notion of one truth is fundamentally divisive; in practice the notion of one truth contradicts the collaborative ethic and supports interaction characterised by entrenched positions. Search for "the" truth is often productive within a closed space of cultural assumption, but does not lead to open inquiry outside that space; rather it invokes defensiveness, and if necessary violence in order to sustain its inviolability. Inevitably it leads to fragmentation and conformity, as contradictory elements break away to form their own "truthful" reality, and all else becomes subservient to "truths" current fashion (Feyerabend, 1988).

One more point about multiple truths; such a claim does not contain the inference of the catastrophic consequence that all "truths," that is, socially acceptable beliefs, are equally useful or sustainable, or that some cannot be falsified. At least at the level of physical definition, it is demonstrably false that I am constructed entirely of green cheese. Such a claim is not a valid contender for any claim to a truth beyond that of a very idiosyncratic and metaphorical form. Truth claims about events can never be proved, but some truth claims can be demolished through procedures of contradiction.

If data belongs to an event, it cannot be attributed to a particular agent or aspect of that event. It is common and comforting to attach data to particular objects or participants in an event, and to the extent that all other participants and relations that constitute the event are held constant and made overt, to that extent attributing the data to a particular agent constitutes a valuable shorthand in description and discourse. For example, to attribute a certain tensile strength to a steel beam is convenient, but has meaning only in regard to an event at which, at a certain temperature, the beam is stretched in a machine until it breaks. The time span within which this (hypothetical) event generates the same data is quite long. But over a thousand years, the steel beam no longer has this property; which is shorthand for saying it will behave differently in the event that it is stretched. Not only that, but any engagement in events will affect the tensile strength in an unpredictable way; if an unbroken part of the beam is stretched again it will be found to have a different tensile strength; as it will after multiple vibrations as part of a bridge.

So experiments in the physical and biological sciences do not produce data about the object, or measure properties of the object being investigated. They produce data about the event that is the experiment. Most experiments describe the behaviour of physical or biological objects under particular boundaried, that is, controlled circumstances. The information they give therefore is not so much about the "natural" world in which we and they live, as it is about the "controlled" world that is the experiment, and sometimes becomes habitualised as technology. Most social research has fallen into this trap of misrepresentation of the source and attribution of data.

Social events, or indeed interactional events of any sort involving living things, have time spans of small duration. Indeed, identical events are impossible to create because social relations, and the participants involved in them, continually change. Even if we could hold all the conditions constant as we do for the steel beam, the data still cannot be attached to the person because, even more so than for the steel, the person of tomorrow is a different person; and part of the difference is attributable to the experience involved in obtaining the data.

It follows from this epistemology that most psychological descriptions of people are shorthand and problematic descriptions of social events, from which most elements that constitute the event are camouflaged. The label is attached to the person even though the events which produced the data involved social interactions. This is an example of faulty labelling. In particular it applies to any notions of skill and competency that do not clearly define the context of their application.

So the issue of objectivity is not that things exist independently of the mind; the issue is whether things (elements) have properties independently of the events used to describe them. To say that a thing is real (has material existence) is very different to claiming that its "properties" are real and belong to it.

Ontology: What is the nature of social reality?

Within the meanings constructed above ontology precedes epistemology in that social relations are a particular case of an event in which two sentient beings (probably both human), are involved. By implication the event is the "reality." Something is happening "out there" that is producing a difference. Thus social experience is a particular form of experience of an event, and social meaning a particular construction of that experience.

On the other hand, epistemology precedes ontology in that all meanings are socially constructed, and are thus ultimately dependent on social relations and that includes the meanings we ascribe to ontology.

Regardless, the two domains interlink with no inconsistency in terms of the idea of social relations and the idea of knowledge being a function of experience of relational events, and meaning being socially constructed.

Using relations as a primary explanatory factor negates the notion of causality, at least in a simplistic sense. Events are construed as interactive systems where everything effects everything else; patterns of mutual influence replace causality as an explanatory principle. This has been generally accepted in Physics since the work of Einstein and Eddington early this century. It has always seemed odd to me that the more complex the system in which the event occurs - from physics through to biology through to social relations - the more frantically the idea of cause is clung to.

Further to that, the idea of "reality" is similar to the idea of "truth"; a redundancy, an unnecessary complexity, an irrelevant diversion. It contributes to conflict rather than to productivity. It seems more useful to talk about what aspects of social relations intrude most on experience, and are important to the intensity and duration of that experience, and the effects that it generates. In this regard I would make four assertions about social events, conclusions from my own experience and reflection:

- knowledge of social relations (that is, data generated within human interactions), is usefully construed in terms of the power and affect relations of the participants in the event; in particular, asymmetrical power relations generate different data than do symmetric power relations; and positive affect different data to negative affect (Foucault, 1988).
- an event occurs within specific localised power and affect contexts; this is not to suggest that this event might not itself be embedded in power relations (economically, racially, nationally or gender influenced) which push the effects and experience of the event in particular directions, but does put less emphasis on such grand power relations.
- events are dynamic, not static situations; they are characterised by movement, by change. They exist in time, which could be considered one measure of their change. So data about social interactions, which may often be characterised by power and affect relations, will change over time as the power and affect relations themselves change. I assume that any new social relationship (any social event characterised by people who have not met before in that configuration) will initially be asymmetric in respect to power, and moot in respect to affect. The relational changes will affect the data generated through interaction, which includes discourse, and vice versa.
- Fixed societal structures (e.g., hierarchies) crystallise power relations and negate change. To the extent that they are successful they may produce knowledge, consensual interpretations, limited by the very boundary conditions that make its production

> possible; fixed societal structures also, in time, contradict the flow of interactional life, and produce social pathology.

Axiology: What values are embedded in the processes and product of the research? Whose interests are served through them?

No knowledge is value free. As Lincoln (1990) puts it, "given the criticism from all quarters, . . . only the most intransigent or the most naive scientist still clings to the idea that inquiry can, or should, be value free"(p82). Being socially constructed, knowledge produced from inquiry is related to the meanings and purposes and structures within which it was composed; and it will tend to confirm or negate those relations involved in its construction, depending on the interests and attitudes and assumptions and awareness of the researcher. Even if data could be produced that was independent of those elements and relations, that very independence is itself a value position, which could be construed either as objectivity, because it has transcended bias, or as ideology, because it camouflages the power relations from which its bias necessarily derives.

As a researcher my task is to contribute to the meaning system that helps me and other people make sense of their experience in the particular class of events with which this study is concerned. They will make sense of it if it is a story that links in some way with their experience, and at the same time is not contradictory to their experience; experience that is, of course, already partly interpreted in terms of other stories.

As an educator my task is to change people; education is nothing if it does not result in change. And as change is inevitable, but may be in many directions, there is obviously an obligation on the part of the educator to specify the direction in which change is intended.

As educator-researcher I must interact with the people with whom I wish to do research or educate. I do this through process (how I do the research), and product (what I produce as a result of the research). If I do not produce the data I investigate, but merely interact with data produced by someone else, this simply pushes the value problem one step backwards; their data was not value free. So if I accept their data without criticism, then I am accepting and perpetuating the values that affected its construction and effects. If I question that data, I question the social values embedded in it, as much as the social effects that are manifested through it.

If whatever I do involves interactions with people, and the construction of knowledge, then whatever I do affects both the meanings of people, and the social relations involved in those meanings. This is not to say that describing "what is" implies approval and acceptance of what is. Rather it is to claim that the very description of "what is" implies a way of viewing the world, a relationship with the situation, an involvement in the construction of the data, that pre-empts the meaning of the data by hiding the value assumptions behind the very mechanisms of its construction; becomes, that is, symbolic violence, unless made explicit (Bourdieu, 1977). Most quantitative research and much qualitative research is in this sense symbolically violent, in that the sources of its power are disguised.

Unless I wish to engage in a value contradiction, it seems necessary to have an awareness of the direction in which I wish to move people's overt and covert experience of social relations and the meaning systems construed within their influence; and to use processes and meanings that are congruent with those purposes.

My autobiographical note indicates that much of my work over the past thirty years has been involved with the nature and practice of violence in its various forms, especially as it affects young people.

My construction of the concept of structural violence (Wilson, 1992) indicates that I regard fixed hierarchical structures, in all their multifarious visible and disguised forms, as inevitably connected to structural violence and hence to social injustice. Due process within legal systems is necessary to alleviate, or control, some of the social fallout, but is not sufficient to ensure social justice at its root manifestation, which requires more equalitarian structures.

Peace and social justice are ideals that have many forms and faces that change over time. On the other hand, physical and structural and emotional and symbolic violence are constructs amenable to more specific definition, and hence more easily recognisable in particular social events. For this reason, I feel more comfortable having as a basic value the reduction of violence, which I could universally advocate, than with the increase of social justice, which is more nebulous because of its many-faceted nature; on this view, increase in social justice that is not associated with reduction in violence would be problematic, involving as it does an internal contradiction.

If beliefs (truths) are multiple, then so must be the values that are implied in those beliefs, or which inform them. How then can any particular value position be maintained as superior to any other?

In regard to the specific events that involve me and others in this thesis, I would answer that while the value of reducing violence is not necessarily superior to others, in the context of this work it is consistent with:

- 1. The learnings (culture and gender influenced as they are) that I have constructed out of my life experiences.
- 2. The ontology and epistemology which I have described, which inform the assumptions on which this study is based.
- 3. A view of life and living that involves ideas of growth, change, and flow at both individual and social levels. As such it is

consistent with many views of personal enlightenment and social justice.

- 4. Processes likely to favour the survival of human life on the planet at a time when the technology is available, and primed to destroy it (Schnell, 1980).
- 5. That universal attunement and compassion which is one aspect of the experience described as mystical, as cosmic consciousness, or as the perennial philosophy, which transcends historical and cultural boundaries, and contains a sense of the sanctity of each individual person (Wilber, 1991).

Slotting into the social research field: How does this epistemology, ontology and axiology fit into the social research field as currently constituted?

Some doyens in the research game still regard qualitative social research as an exotic rather than a native plant, and as such something to be treated with caution because of its possible ecological effects on what had previously seemed to be a very secure and threat-free environment. Specifically, many testing experts still live in a positivist world (Shepard, 1991). As well, most teachers are quite convinced that their tests measure their student's attainments; the correspondence theory of knowledge may well be discredited, and philosophically empiricism may well have been dead for forty years (Smith, 1993), but in schools and colleges and universities and work places it is alive and kicking. However, a rich literature has developed from the debates involving qualitative research over the last ten years (Burgess, 1985; Eisner & Peshkin, 1990; Guba, 1990 Popkewitz, 1984; & Smyth,1994).

So with some reservations qualitative research is now accepted and respectable, even though practice severely lags theory. The reservations are currently crystallising as sets of questions and answers about how to recognise "good" qualitative research. For example Carr and Kemmis (1985) describe five formal requirements for any adequate and coherent educational science (p158). Criteria and caveats are being constructed that will undoubtedly in time result in a new orthodoxy (Lincoln, 1990). Feyerabend's (1988) assertion that "science is an essentially anarchic enterprise; theoretical anarchism is more humanitarian and more likely to encourage progress than its law-and-order alternatives"(p5), provides as much discomfort in the research world, be it quantitative or qualitative, as in the world of politics or the family. Smith's (1993) work clearly indicates that clarification of the problem of criteria is central to any real progress. It is also necessary if any substantial change in educational practice, and associated structural relations, is to occur.

At this point in time, however, the limits of the field are blurry, and the demarcations between various camps subject to border skirmishes. So at least one reason for my position not fitting into a specific ontological, epistemological, axiological, or methodological tent is that such tents are not clearly differentiated between the encampments. Having said that, it is possible to nominate some camps to which I do not belong, and some camps to which I partly belong, where I would not feel too uneasy sitting in some of their tents.

It is generally agreed that there are three basic positions; empiricist (post positivist), interpretivist (constructivist), and criticalist (Smith, 1994; Lincoln, 1990). It is also agreed that this is an over simplification.

Briefly, empiricists argue that there is a reality out there to be discovered, that it is single and measurable, and that causal laws explain and predict it (Smith, 1994).

Carr and Kemmis (1983) characterise the interpretive approach to social science as aiming "to uncover the meaning and significance of actions" (p92). The interpretive position is that truth is constructed by people, and always involves a social context and social interactions. So truth is relative and multiple. This position has two strands, the ethnographic (Sherman & Webb, 1988), and the ontological strand (Eisner, 1988). The difference is in the way hermeneutics is regarded. In the ethnographic strand, hermeneutics is a method of achieving interpretive explanation; in the ontological strand hermeneutics is more concerned with the idea that all knowledge, all representation is dependent on the primacy of experience (Schwandt, 1990). Regardless, "hermeneuticists of all measure and variety agree that any interpretation of meaning must take place within a context" (Smith, 1993, p16).

Carr & Kemmis (1983) regard post-positivist and interpretivist accounts to be similar in that "the researcher stands outside the research situation adopting a disinterested stance in which any explicit concern with critically evaluating and changing the educational realities being analysed is rejected"(p98). However, some constructivists (Lincoln, 1990), more recently advocate an abandonment of "the role of the dispassionate observer in favour of the role of the passionate participant" (p86). This is a position with which I concur. Smith (1993) elucidates other similarities and differences in the various positions:

> Interpretivists take antifoundationalism to mean various closely related things such as that there is no particular right or correct path to knowledge, no special method that automatically leads to intellectual progress, no instant rationality, and no certitude of knowledge claims. These are ideas, of course that interpretivists share at one level or another with postempiricists and critical theorists (p120).

He goes on to point out that "differences of consequences are readily apparent as these points are elaborated upon more specifically"(p120), and presents his own view that

> the demise of empiricism means that it is time to move beyond the need for a theory of knowledge and the various dichotomies . . . of subject versus object, facts versus values . . . this is in marked contrast to attempts by post empiricists and critical theorists to elaborate a successor theory of knowledge by either modifying or recasting, respectively, the empiricist understanding of these dichotomies (p120).

The criticalist position also has two strands. In the first belong critical social theorists, ranging from traditional Marxists uncovering the "contradictions of economic conditions and relationships", to a variety of other critical perspectives, where "the focus is on the ideological distortions inherent in a broad range of historically formed social and cultural conditions" (Marshall, 1990, p181). Smith (1990) sums up the critical theorists project: "critical inquiry can reveal our objective historical conditions: tie this knowledge to the expunging of false consciousness, distorted communication, and so on; and thereby promote emancipation and empowerment" (p193). Critical theorists then have a clear agenda of social transformation, based on a particular historical perspective, to which they have appropriated the "objective" label. As Carr and Kemmis (1983) express it, they aim to "reawaken the power of criticism and the power of praxis - criticism and praxis being the critically enlivened forms of what we usually refer to as theory and practice" (p186).

The other strand of the criticalist position is the post-structural, post-modern strand, which includes some feminist perspectives. The concentration here is on the construction of social reality through language and discourse, and the way in which this serves dominant groups and interests. The emphasis in research is on discourse analysis, in order to expose such inequities (Smith, 1994). Foucault's work is sometimes attached to this strand, though he himself did not accept the classification. And I would agree. This is important, because the writings of Foucault considerably influenced this study.

So where does my position fit into all this? I am not a positivist or empiricist. I do believe that empirical data can be collected about events; it's just that I don't believe that in relation to social events such data is very stable, can be replicated without considerable error becoming evident, or can be justifiably attached to a particular participant constituting the event. Any such data views that event from a particular position, with particular boundaries, with particular interests and values influencing the collector.

On the other hand truth claims are sometimes explicit, and often implicit, in theoretical formulations or interpretations involving social events. And some such claims can be directly contradicted by empirical data, by effects or consequences that are directly observable.

In terms of ontology, of the nature of reality, I do not fit neatly into any of the camps; empiricist, interpretivist or critical. I am probably closer to being a sceptical mystic. Rather than enter into that potential bog, in this thesis I have bypassed the question of "reality" and begun with the notion of social events, which involve the participants in social experiences.

I am constructivist or interpretivist in as much as I see all knowledge as multiple and constructed. Eisner (1990) agrees that experiences are the basis for cognition and knowledge: "thinking and knowing are mediated by any kind of experiential content the senses generate...our language refers to referents we are able to experience, recall or imagine"(p91). However, as Schwandt (1990) points out, this ontological basis of experience is not common to all interpretivist methodologies.

Perhaps my main point of departure from the criticalist perspective is at the ontological level; certainly I see relations as fundamental in as much as they constitute the mechanisms through which difference and change occur, thus making events experiencable. But I do not wish to "objectify" these into some grand historical schema on the one hand, nor overemphasise their dependence on gender relations or particular discourses on the other. Rather, I see power and affect relations as a "heuristic fiction" that has great generality and elegance as an explanatory and generating principle. However, I am clearly allied with them in their wish to reduce the violation of persons through the transformation of social structures and in seeing social research as a legitimate way to help people make sense of the social world in a way that gives them some leverage to change it for the better. By "better" I refer to a decrease in violence.

A model for the assessment process

This thesis is concerned with a particular type of social event called assessment. It is particularly concerned with the assessment of individual persons. I assume that such an assessment results in a categorisation of some kind. Such a categorisation involves a bifurcation of data, itself dependent on judgments about criteria and standards.

Given the ontological position of the above discussion, the assessment process involves (at least) five stages (events) and a context. In actual practice some of these stages may be omitted or fused. Such fusion or omissions may constitute a source of confusion or error.

- 1. Test production: An event (experiment, test) is devised to produce data. Such an event will involve an interaction between the assessed person, and instrumentation of some kind. The instrument may exist in the assessor's head, or may be produced as a physical artifact (a written test). The test production process also involves explication of a theory-practice link of some sort, and some prior judgments about a relevant task.

- 2. Test experiment: The person being assessed does the test, by performing what is required in the testing situation. This is the first stage of data production, and this event is completed when the test is completed.
- 3. Data production: The second stage of data construction occurs when the assessor interacts with the testing process directly, or with products from it. eg. a performance or a completed test paper. This interaction involves an interpretation of the data.
- 4. Judgment process: This results in a categorisation of some kind; it involves a comparison of the data with the standard, either directly, or by comparing with data about other students. This process assumes the existence of the standard as a stable and replicable element in the event.
- 5. Labelling process: At least two labels are involved; the name of what has been assessed (described), and the name that describes the level of performance (compared to the standard). The multiple label is constructed from the whole assessment process, and is legitimately attached to those events. In practice it is more likely to be attached to an element of the testing event (the assessed), or to an even more remote theoretical construction related to the assessed (some skill or ability).
- 6. All of these processes are embedded in relations of power which reproduce and invigorate themselves in the processes. And all of these processes (events) are potential sources of error and confusion in the individualised material product of this whole process - the documented labelling and categorisation of the assessed person.

Summing up

Negating notions of truth and reality does not necessarily lead to chaos or alienation, but may presage a search for greater clarity of assumption, for greater precision of value, and hence for greater wisdom in action.

Return to Table of Contents

Part 2: Context

Chapter 4: Power relations
Chapter 5: Power relations in educational systems
Chapter 6: Standards, myth and ideology

Chapter 4: Power Relations

Synopsis

Power is defined in terms of relational fields rather than of personal or role attributes, of power as ruler and ruled. Arendt and Foucault articulate the construct differently in that they differentiate violence from power. I choose a broad definition of violence as any violation of personhood; so both force and physical violence are subsumed as sub-categories of that construct; and violence becomes a necessary aspect of asymmetric power relations, inevitable in hierarchies.

The other side of power relations is now highlighted; the side that produces rather than denies, that constructs rather than destroys. That is, I deal in some depth with Foucault's (1992) assertion that "power produces; it produces reality; it produces domains of objects and rituals of truth. The individual and the knowledge that may be gained of him belongs to this production"(p194). In particular, I look in detail at what is produced through two specific mechanisms fabricated within asymmetric power relations: the processes of disciplinary power, regulated through surveillance and penalty; and normalisation, achieved through linear labelling and sustained through the cult of individualism.

I look briefly at some of the "scientific" disciplines, and the micro-cultures that sustained them and helped provide their assumptions, theories and data.

Finally in this section Bourdieu's construct of symbolic violence, and the notion of habitus through which it is humanly experienced, shows how difficult it is, when playing the game our culture dictates, to recognise its limitations.

Defining power

What characterises social life is affect and effect; affect refers to those aspects of relating that are characterised by polarities such as emotional closeness-distance, of like-dislike, of attraction-repulsion, of affiliation-separateness. These affect relations are apprehended viscerally, experienced directly through the body. In the vernacular, in the field of sense relations you "feel the vibes."

Power refers to those aspects of relating that translate influence, that make a difference, that have an effect. The actions of one affect the thoughts or actions of another. The poles of a power relation could be characterised by such descriptions as dominant-submissive, controlling - rebellious, have - want, strong - weak. So within the field of power relations, what one person does affects a second, which affects a third, and so on. Such effects ripple onwards and outwards from human interactions in patterns that are indeterminate; yet even so the patterns are sometimes decipherable and probablistically predictable, for the fields that affect the patterns are stable and translatable.

For example, in all cultures there are families, groups of people genetically related whose patterns of interaction are relatively stable, whose ways of behaving towards one another are consistently patterned; the parent influences the child, the parent's demands produce action, the power vector is from parent to child. Yet even so the child's behaviour must influence the parent's behaviour, if only to maintain the parent's controlling function. In this sense power relations involve mutual influence, even though normally asymmetric, and translated into action involve dynamic events.

Such events are acted out in power fields, such as family or school or workplace, where the rules of the game are understood, and the overall direction of action influence predictable. In this sense the influence is not so much person to person as role to role; the relationship of parent to child overrides the relation of the person Jack to the younger person Julie. For this to occur we must assume some mechanism for the learning of relational roles, for the internalisation of the power injunction. For if we locate the power in a relational vector out there in the space between, we must also explain by what psycho-social means people in the field are moved to act. More of this later.

Affect and power relations are not mutually exclusive; strong affect can generate high intensity in the field of power relations. And doubtless asymmetric power fields are capable of generating considerable affect, both positive and negative. Even so, the two notions are separate, the two fields initiate different experiential effects, and are associated with different states of consciousness. Love and power are not synonymous. And which is stronger is moot. Like Bourdieu (1990 a), "We leave it to others to decide whether the relations between power relations and sense relations are, in the last analysis, sense relations or power relations"(p15).

Regardless of their relative strengths, their confusion produces dysfunction in societal relations, and pathology in individual people; love that degenerates into power play destroys itself; and power that masquerades as love is a sickening violation. However, this is too large a contention to debate in this thesis, and is not directly related to our major theme (Laing, 1967).

To summarise, I have defined power relations as the dynamics of mutual influence. In most situations such relations are activated in fields whose pattern is perceived by those who enter the field in terms of role relationships, or less consciously simply as appropriate behaviour, a predisposition to act in a certain way. People engaged in such fields are both activated and constrained, but by no means wholly determined, by the role expectations or predispositions (habitus) which, for individuals at either pole of a power relation, are activated by their entry into the field.

So let's see how this definition fits into the historical meaning of such concepts as power, force, strength, and violence.

Power and Rule

Traditionally the essence of power has been rule and command; or alternatively the act of ruling and commanding has been attributed to a faculty called power. This need to dominate was seen as an instinct in man, a psychological necessity. Force and violence in social life was thus inevitable, for they were necessary components in the command strategies of a leader. Combine this psychological instinct with the social requirement that the first learning of civilisation is that of obedience, and the two poles of a largely unidirectional power relation are accounted for. To command and be obeyed is thus the essence of Power. And the basic building block for monarchy, hierarchy, and their complex transformations into the modern state has been constructed (Arendt, 1970, p36).

A look at any parliament in action, or a peep into any political party meeting, leaves little doubt that this paradigm of the fight for dominance is still central to the inner workings of government; certainly jostling for place in the political party pecking order is a major preoccupation of politicians, particularly of those who aspire to top positions. However, tradition also specifies an alternative power game.

This was the idea of representative government, where obedience is to laws that have the people's consent rather than to dominant men, and elected leaders remain dominant only with the support of the people. This second paradigm undoubtedly has a much wider gap between vision and practice than does the first, and a fundamental question of political science has always been about whether this is ideology rather than reality, a fairy story that disguises and soothes the experience of most people of powerlessness, of alienation. Regardless, in most modern states there is some balance, some checks within limits, of the power of the state and the tyranny of its accompanying bureaucracy, articulated through the opinion of the people.

Arendt (1970) argues that all government - tyrannical, monarchical, oligarchical, democratic, bureaucratic, or whatever, depends finally on the support, the "qualified" obedience, of the people:

> All political institutions are manifestations and materializations of power; they petrify and decay as soon as the living power of the people ceases to uphold them. . . (so) one of the most obvious distinctions between power and violence is that power always stands in need of numbers, whereas violence up to a point can manage without them because it relies on instruments (p41).

Arendt wants the word power to be reserved for the many, as distinct from strength, which is a property of the singular, a function of character or charisma or physical prowess. So an individual who appears to have power has it only in relayed form from the many whose support is needed. Whereas violence uses implements to multiply strength.

Power and structures

What characterises all of these notions of power is their attachment to particular agents, either singly or in groups. Power is a quality, a property, of an object or objects. But there is another way of viewing power:

> The major contribution of what one has to call the structuralist revolution consisted in applying to the social world a relational way of thinking, which is that of modern physics and mathematics, and which identifies the real not with substances but with relations (Bourdieu, 1990 b, p126).

Bourdieu postulates the existence in the social world of objective structures, in addition to symbolic systems, and independent of consciousness and desires of agents; structures which guide and constrain their practices and representations, which produce a predisposition to act in certain ways (p123).

Foucault (1988) also moves well beyond the notion of "Power - with a capital P - dominating and imposing its rationality upon the totality of the social body." In fact, Foucault goes on to say, "there are power relations. They are multiple; they have different forms, they can be in play in family relations, or within an institution, or an administration - or between a dominating and a dominated class" (p38).

Foucault (1988), like Bourdieu, uses the relational power structure as a fundamental explanatory principle: "The characteristic of power relations is that, as agents in the structure, some men can more or less determine other men's conduct, but never exhaustively"(p83). So power relations precipitate all "the strategies, the networks, the mechanisms, all those techniques by which a decision is accepted and by which that decision could not but be taken in the way it was"(p103). Or in retrospect, that's the way it seems.

Power and violence

Yet like Arendt, Foucault (1988) wants to remove coercion, brute force, from his notion of power relations. He says:

> A man who is chained up and beaten is subject to force being exerted over him. Not power. But if he can be induced to speak, when his ultimate recourse could have been to hold his tongue, preferring death, then he has been caused to behave in a certain way. His freedom has been subjected to power. He has been submitted to government. There is no power without potential refusal or revolt (p83).

Yet the man chained does have a choice; to scream or not to scream. And surely Foucault would himself argue that what is conceived as an "ultimate resource" is itself a social construction - more a production of the particularities of his cultural experience than of some "essence" of humanness. And if so the difference he postulates dissolves.

Foucault (1982b) insists that

> What defines a relationship of power is that it is a mode of action which does not act directly or immediately on others. Instead it acts upon their actions: an action upon an action, on existing actions or on those that may arise in the present or the future. A relationship of violence acts upon a body or upon things; it forces, it bends . . . A power relation (demands) . . . the one over whom power be exercised be thoroughly recognised and maintained to the very end as a person who acts: . . (so that) a whole field of responses, reactions, results, and possible inventions may open up (p220).

In an otherwise articulate and logical essay on The Subject and the Power written at the end of his long career, Foucault in this passage seems to get lost. Actions now act directly on indefinite actions in an indefinite future in utterly magical ways; if power acts on the body it doesn't act on an action; the person at the dominated end of the power relation has to be recognised. By whom? Most of this is contradictory to all those subtle and unconscious "strategies, networks and mechanisms" through which he says the effects of power structures are promulgated.

There is some romantic idealism involved in this refusal to see violence as a special case of power relations, in this wish to make it a separate category. As Arendt (1970) admits, "nothing . . . is more common than the combination of violence and power, nothing less frequent than to find them in their pure and therefore extreme form" (p46). So what, if anything, is gained by making of violence a separate class of event? Is it that to separate them is to separate the human body, which can be subjected to the ravages of violence, from the "human spirit", which relates to power and can remain inviolate? This is a separation deeply ingrained in Western culture, which denies the integrity of the human organism, and wishes to separate body from soul, and nature (which includes woman) from man.

Perhaps both Foucault and Arendt, appreciating the necessity of power relations for all social functioning, and wanting to emphasise its positive constructive side, want to remove from its definition that which utterly negates the possibility of a spirited response; want to leave open the possibility of a political response in asymmetric power structures that are aided by overwhelming instruments of violence.

In other words, they reject a notion of structuralism in which only surfaces of humans, their bodies and behaviours, are involved; they wish to include the spirit, the internal meanings, as part of the equation; and the confusion arises from their own lack of clarity about how to slot in the subjective element.

Regardless, if we refuse to reify violence, and see it as a process, an interaction in which a living being is violated, then it becomes impossible to separate power relations and physical violations in this way, and it is clear that violations of an instrumental kind are but one strategy in a whole armoury of mechanisms available in the field of power relations for violating people.

Violation of personhood

Brown (1973) encapsulates this view in his definition of violence:

> The basic definition of violence (is) violation of personhood . . . And since personhood means the totality of the individual, and never just the body or just the soul, we are reinforced in our notion that violation of personhood can take place even when no overt physical harm is being done. In the broadest terms then, an act that depersonalizes would then be an act of violence, since . . .it transforms a person into a thing (p1).

So abuse, beatings, injury, torture and killing, what we normally recognise as violence, are more obvious forms of violation, and perhaps it is the intention to harm and the personalization of the act that makes such actions so abhorrent; the killing of a child with a bayonet seems more heinous than the more objectifiable destruction of a city with bombs. There is a different focus. Yet in the sum total of human misery and violation such intentional physical violence is minuscule.

People certainly are violated when abused or beaten or injured; yet just as certainly are they violated when disregarded or denied, infringed upon or intimidated.

People are disregarded when they are denied the basic rights of food, shelter or care, or full human status in communities. The mechanics of this disregard may be articulated through many systems, based on economics, class, caste, colour, gender, ethnicity, age, religion, or whatever; or more often some combination of these.

Denial, not recognising their existence as fully human persons, is one of the cruellest ways of violating, especially when perpetrated on young children, with its ultimate internalization of the destructive self image "I don't exist."

At a more general level, any positivist stance that treats people as objects, that directly or indirectly ignores of depreciates the internal meanings people create of events, is a violation of their personhood. On this basis much of current political ideology, economics, sociology, psychology, psychiatry, medicine, and educational and management practice, must stand condemned.

People are infringed upon in many ways: police or media or sexual harassment, smoke pollution in public places; confinement in school classrooms. Emotional or symbolic infringement is more subtle: a mother withdrawing love for disobedience; a preacher selling eternal insurance through inclusion in a particular group.

Intimidation also takes many forms; at its most obvious it is the threat of physical pain, at its more subtle the threat of hell. Intimidation feeds on fear; its father is the sword, its mother the imagination. Civilisation enshrines it in Law.

For the more sophisticated, intimidation is predicated on shame and guilt. Shame is the internalization of society's adverse verdict on behaviour, self disgust generated by what others think. Guilt represents a deeper internalization, the adverse criticism of self by self. Of all forms of human violation, the inculcation of guilt is perhaps the most oppressive, for guilt is pervasive in its influence and insidious in its effects.

In addition, humans are growing organisms. Their normal state is development, not stasis. So humans are violated not only when their physical existence or their psyche is threatened, but also when their capacity for growth is stunted, when their potential for expansion is diminished (Wilson, 1991, p16).

So we approach a dilemma: power structures are cultural necessities, the essence of community life, and at this point in cultural history all cultures are predicated in one form or another on asymmetric power relations; and all of the violations described above are manifestations of asymmetric power structures. It follows that violence necessarily flows from human culture as currently experienced. And attempts to separate power from violence involve inherent contradictions.

Power and production

One issue here is not whether asymmetric power relations predispose violations. They do. An equally important issue is whether they also have a productive role to play in the human condition. And they do. Foucault's great contribution has been to spell this out. "The refusal, the prohibition, far from being essential forms of power, are only its limits, power in its frustrated or extreme forms. The relations of power are, above all, productive" (Foucault, 1988, p118).

This view does redress the balance and help us to see the other side of the coin. People are produced and reproduced through their immersion in power structures. So are cultures. And the human spirit sometimes soars above the violence. Even so, the violations are often not extreme forms; they are inherently, pervasively and insidiously embedded into the structure.

So we must ask, what does "productive" mean in this context? If knowledge and people are socially constructed, what constitute the productive, rather than destructive manifestations of power relations? From what frame of reference is the separation between intellectual or emotional production and destruction recognised? As a starting point, let's first look briefly at Foucault's views about the mechanisms of this production, and then at Bourdieu's ideas about the inevitability of symbolic violence within reproductive cultures.

Disciplinary power

Over the past three hundred years, power on this planet has assumed a new face. Foucault (1992) traces this transformation brilliantly in Discipline and Punish:

> Traditionally, power was what was seen, what was shown and what was manifested, and paradoxically, found the principle of its force in the movement by which it deployed that force. Those on whom it was exercised could remain in the shade; they received light only from that portion of power that was conceded to them, or from the reflection of it that for a moment they carried. Disciplinary power, on the other hand , is exercised through its invisibility; at the same time it imposes on those whom it subjects a principle of compulsory visibility . . . the examination is the technique by which power, instead of emitting the signs of its potency, instead of imposing its mark on the subjects, holds them in a mechanism of objectification (p187).

Foucault is using the term "examination" here in its widest context. The written test as we know it is a refined and intense form of that "hierarchical observation" and "normalizing judgment" that characterise all examinations, whether they be pedagogic, medical, legal, penal, supervisory, psychiatric or whatever.

How is this power transmitted? What is the mechanism of its distribution?

> The power in the hierarchized surveillance of the disciplines is not possessed as a thing, or transferred as a property; it functions like a piece of machinery. And although it is true that its pyramidal organization gives it a "head," it is the apparatus as a whole that produces "power," and distributes individuals in this permanent and continuous field. This enables the disciplinary power to be both absolutely indiscreet, because it is everywhere and always alert, since by its very principle it leaves no zone or shade and constantly supervises the very individuals who are entrusted with the task of supervising; and absolutely "discreet," for it functions permanently and largely in silence. Discipline makes possible the operation of a relational power that sustains itself by its own mechanism and which, for the spectacle of public events, substitutes the uninterrupted play of calculated gazes (p177).

The details of this disciplinary power seem trivial in their manifestation:

> The workshop, the school, the army were subject to a whole micropenalty of time (lateness, absences, interruptions of tasks), of activity (inattention, negligence, lack of zeal), of behaviour (impoliteness, disobedience), of speech (idle chatter, insolence), of the body ("incorrect" attitudes, irregular gestures, lack of cleanliness) of sexuality (impurity, indecency). At the same time, by way of punishment, a whole series of subtle procedures was used, from light physical punishment to minor deprivations and petty humiliations (p178).

Together these trivialities articulate a milieu, produce an enveloping social environment, so that the people who live in that space accept it as a way of life, as a natural way of being. And so we find that, in the field of education

> A relation of surveillance, defined and regulated, is inscribed at the heart of the practice of teaching, not as an additional or adjacent part, but as a mechanism that is inherent to it and which increases its efficiency (p176).

Praise and blame

Disciplinary power uses the twin instruments of observation and judgment, and the judgment is by necessity judgmental; is categorised by a satisfactory- unsatisfactory dichotomy. Such normalizing judgments are so pervasive as to override their specific instances. "Humanistic" teachers may protest that they punish the misbehaviour and not the person; this may be true of their intentions, but does not describe the effects. Again Foucault spells it out; the judgments not only diminish the aberrant behaviour; they also produce the person:

> Through this micro-economy of perpetual penalty operates a differentiation that is not one of acts, but of individuals themselves, of their nature, their potentialities, their level or their value. By assessing with precision, discipline judges individuals "in truth"; the penalty that it implements is integrated into the cycle of knowledge of individuals (p181).

This translation of act into essence, of misbehaviour into attitude, of error into ignorance, of absence into inability, is one of the political functions of Psychology. This transformation of event into label is an epistemological error, a misrepresentation of the functioning process, but is crucial to the construction of those "individuals" of whom Foucault speaks. For as he indicates so clearly, that individual first constructed in the eighteenth century, that educated individual being continuously recreated in "developed" twentieth century countries, is not characterised by passion, creativity and an independent mind. On the contrary, the individual is a person cleverly moulded by disciplinary power to be utterly reasonable (that is, to deny emotion), completely responsible (that is, to deny spontaneity and creativity), and to be loyal and dependable (that is, to deny independent thought and action).

Illich (1971) reached similar conclusions:

> Under the authoritative eye of the teacher, several orders of value collapse into one. The distinctions between morality, legality and personal worth are blurred and eventually eliminated. Each transgression is made to be felt as a multiple case. The offender is expected to feel that he has broken a rule, that he has behaved immorally, and that he has let himself down (p32).

Normalizing

This process of creating the conformist and at the same time supporting the cult of the individual, is what Foucault calls normalizing. It involves five distinct operations. "The perpetual penalty that traverses all points and supervises every instant in the disciplinary institutions compares, differentiates, hierarchizes, homogenizes, excludes. In short, it normalizes" (Foucault, 1992, p183).

So what a child (or adult) does is seen not in its own right, but in the light of what others do. Behaviour and product, and ultimately relations and being, are constructed and thus perceived and conceived in comparative terms. So I do not exist in relation to others, but in comparison to them; I become an object in the field of comparison, rather than a subject in the field of creative and responsive relation.

The thrust of this comparison is not identification, but differentiation; the comparison focuses not on the similarities, but on the differences. The effect then is not to produce belonging and cohesion, but rather alienation and separation. And this differentiation is not in terms of the infinite variety of human behaviour and persona, but within a simple hierarchical catagorization of better or worse. To achieve this it is necessary to collapse the variety, the complexity, into a few single dimensions of value. And because the individual performances are indeed always multi-dimensional, and idiosyncrasies always do become visible, it becomes logically necessary to attach the value to the person, and not to the performance. The notions of skill, ability, attitude, intelligence, competence, morality, are uni-dimensional, and thus can be categorised and hierarchized as more or less, because they meet the joint requirements of unity and invisibility, and incidentally, of fantasy. (This argument is developed more fully in the chapter on comparability.)

And so we become homogenised, perceiving ourselves, and thus being ourselves, in the times and places constructed for us along the one-dimensional spaces into which we are constrained. It is as though hundreds of cakes, all made of different quantities of different ingredients, have to be rated in a competition. It is noted that most of the cakes expand on cooking. So we create a single variable called sponginess as a major dimension of comparison. Now we can proceed. The cakes are all more or less spongy. Now comes the moral shift. Some, indeed, are seen to be too spongy or not spongy enough. And so there evolves a notion of value within limits, of quality defined by conformity, of a homogeneity to which all good cakes must aspire.

These processes of comparison, differentiation and hierarchization lead necessarily to notions of the normal, of the acceptable, to the limits within which life must be lived, and outside of which punishments naturally accrue. The pervasive threat and final punishment is exclusion.

These modes of living are learned in most family settings, but the school classroom is the great levelling field where it pervades the life of the group. It is this pervasive quality that so affects the way of seeing other people and oneself that any other way seems alien.

In the late 1970s I was involved in a project in secondary schools involving non-judgmental assessment of students. That is, assessments that simply stated what they had done without that statement containing overtones of satisfactory-unsatisfactory, good-bad.

We explained to over a hundred teachers what we wanted. We asked them to consider particular students whose work they knew well, and to describe some particular examples of their work in this way. We ended up with some two hundred descriptions, of which we hoped to use twenty in our report as examples of non-judgmental descriptions of student work. In fact, none of them was suitable. The teachers were simply unable to write such descriptions; they were unable to see their students (or their student's work) in other than normalizing terms.

Their reality, based on standards, nullified their best intentions.

Individualism

We must not confuse the individualism of our current society with that myth of wild west rugged individualism which is part of the American dream, and exemplifies the "Aussie battler," though doubtless ideologues might welcome the confusion. The individual differences we produce are characterised by creating levels within homogeneous orders, by categorising along linear dimensions of value, by dichotomising continuous performances.

The person's individuality is thus produced by placing him or her along a simple scale, good or bad, satisfactory or unsatisfactory, suitability or unsuitability along a number of dimensions. The individual becomes categorised, described, and indeed produced by the grade, the mark, and finally the profile, which becomes the true description of the shape of the person.

The disciplines

Before we look in more detail at how the formal examination fits into all this, and more specifically the part that the notion of standard has to play, it is useful to fit this development into an historical context. For life was not always this way:

> Historically, the process by which the bourgeoisie became in the course of the eighteenth century the politically dominant class was masked by the establishment of an explicit, coded and formally egalitarian juridical framework, made possible by the organization of a parliamentary, representative regime. But the development and generalization of disciplinary mechanisms constituted the other, dark side of these processes. The general juridical form that guaranteed a system of rights that were egalitarian in principle was supported by these tiny, everyday, physical mechanisms, by all these systems of micropower that are essentially non-egalitarian and asymmetric that we call the disciplines. And although, in a formal way, the representative regime makes it possible, directly or indirectly, with or without relays, for the will of all to form the fundamental authority of sovereignty, the disciplines provide, at the base, a guarantee of the submission of forces and bodies. The real, corporal disciplines constituted the foundation of the formal, juridical liberties. The contract may have been regarded as the ideal foundation of law and political power; . . . The "enlightenment," which discovered the liberties, also invented the disciplines (Foucault, 1992, p222).

Here then, brilliantly summarised, is the monstrous double bind that accompanied the introduction of parliamentary democracy, the genesis of that sense that all thinking people have of "with all these freedoms, how come I don't feel free?" And looking around, they do see all those economic, class, race, gender sources of inequality, and direct their attention to their amelioration, and forget that all were constructed out of the same structural cake mix, from the relations of disciplinary power embedded in hierarchy.

Yet there was a further development here that added immensely to the effects. The hospital, the school, and the workplace, once they had become located as gardens for the growth of disciplinary techniques, at the same time provided nourishment for the accumulation of new branches of knowledge. Clinical Medicine and Psychiatry became branches of knowledge predicated on hospitals and asylums; Education and Child Psychology were branches of knowledge predicated on schools; and Management Theory is predicated on offices and factories. (Offices are no less offices because their power relations and communications are crystallised through computers and their agents can be physically widely dispersed).

It is important to realise that these branches of knowledge developed after the structures, both physical and relational, were in place, and not the other way around. What we have here is knowledge developed within institutionalised relations; knowledge of people already objectified by disciplinary power; knowledge, that is, predicated on institutional inequity, and thus committed to rationalising that objectification.

So pedagogy is knowledge of the learning of children confined in classrooms, just as child developmental psychology is an accurate description of the growth patterns of children produced (both constructed and oppressed) in family and school. When the common translates into the normal and hence the real, these descriptive charactertures define the nature of children.

The unexamined givens of these systems of knowledge are the institutions in which they are based, just as the power relations that are embedded in these institutions comprise the assumptions on which these disciplines are built. And in its turn, the knowledge produces a magnification of that power asymmetry, both because it forms the basis of a verbalised truth that necessarily supports the institutional structure, and because it becomes the property of the professionals who practice it, thus necessarily excluding all others from its mysteries.

Ideologically, these disciplines claim to modify the negative effects of disciplinary power, which

> .seems to have undergone a speculative purification by integrating itself with such sciences as psychology and psychiatry. And, in effect, its appearance in the form of tests, interviews, interrogations and consultations is

> apparently in order to rectify the mechanisms of discipline: educational psychology is supposed to correct the rigours of the school, just as the medical or psychiatric interview is supposed to rectify the effects of the discipline of work. But we must not be misled; these techniques merely refer individuals from one disciplinary authority to another, and they reproduce, in a concentrated or formalized form, the schema of powerknowledge proper to each discipline . . .the examination . . .is still caught up in disciplinary technology (Foucault, 1992, p226).

Now perhaps we can begin to get a little glimpse at the forces that we are contending with here in the field of education. If Foucault is right, then the tenacity of the examination as an educational technique, no matter how professionally denigrated, is easier to understand. And if, as I shall try to show, the examination has no teeth, indeed becomes a paper tiger, without the notion of the standard to support it, then we begin to understand why the empirical facts about the instability, idiosyncrasy, non-transferability - in short, the factual non-existence - of the standard and its measure, has been so consistently and successfully suppressed and repressed.

In the following passage Foucault (1992) indicates the centrality of the idea of the standard. And whilst he is referring here more to standards of social behaviour, they apply equally to more cognitive matters:

> in the genealogy of modern society, they (the minute disciplines) have been, with the class domination which traverses it, the political counterpart of the juridical norms according to which power was redistributed. Hence, no doubt, the importance that has been given for so long to the small techniques of discipline, to those apparently insignificant tricks that have been invented, and even to those "sciences" that give it a respectable face; hence the fear of abandoning them if one cannot find any substitute; hence the affirmation that they are at the very foundation of society, and an element in its equilibrium, whereas they are a series of mechanisms for unbalancing power relations definitively and everywhere; hence the persistence in regarding them as the humble, but concrete form of every morality, whereas they are a set of physico-political techniques (p223).

Educators have been slow to appreciate the implications of Foucault's work to their own discipline. Foucault and Education (Ball, 1990) does explore this domain. And many of the contributors to this book identify the examination as the crucial stategy for embedding knowledge relations into power relations. For example, Hoskin (p31-32) and Jones (p84-97) identify the examination as the pivot of those small techniques through which the modern person is both constructed and controlled.

Symbolic Violence

Before discussing further the place that the examination plays in disciplinary power, I want to examine in more detail the notion of symbolic violence, and the particular way in which it is concerned in the continuance and intensification of violating structures through the imposition of meanings.

The child who is beaten by her father, and is then told that it is God's command that she must always love and respect her parents as indeed her parents love and respect her, and whatever they do is for her own good, is being subjected to symbolic, as well as physical violence. Her experience of being violated is being contradicted and negated. She is told that she is not being violated, but is being helped and loved. And it is not her parents who wish this, but God. She is unable to see that the perpetrators of the violence, and of the meaning system, are both primarily concerned to maintain their own, and each other's, authority structures; that is, the hierarchical power structures that have become institutionalised as family and church. And it is the institutions themselves, not parental love or god, that legitimise the violence, and the justification for it. So these structures become stronger, and the human victims more confused and powerless.

Let's take another example from schooling. Some young people are denied the right to continue their studies. Schools deny them access to further education and hence exclude them from a number of occupations. This is obviously a violation and unjust, even before we look at the inequalities of exclusion in terms of social class, gender and race. How is this exclusion achieved? Schools impose what specific knowledge and skills will be taught, and in so doing define what is useful and legitimate knowledge, and how it will be taught, learnt and assessed. And these processes discriminate against certain groups, and certain particular sorts of people.

The exclusions are legitimated supposedly through the professional judgment of the teacher, who is able to distinguish a "pass" from a "failure." In fact, this is not true. It is the institution itself, the school, that legitimises the exclusion, and inclusion. For the teacher outside the institution, no matter how highly qualified professionally, cannot accredit. On the other hand, the institution can accredit with a multiple-choice, computer-marked assessment system that completely bypasses the professional teacher. So what are in fact rather arbitrary impositions by the school are disguised as professional judgments about skill, ability, and intelligence, and then codified pass or fail with the appropriate label attached to the student. These judgments are then accepted as legitimate by all parties involved, including the great bulk of excluded students, who know at one level that they have been duped, but don't know how.

In these two examples I have tried to elucidate the particular properties of symbolically violent meanings. Firstly they are meanings imposed and legitimated by institutions of authority. For example, by institutions that control morals or education or health or information. Secondly they are designed to convince that what is violent is indeed not so. That what is unjust is indeed just. That what is inequitable is indeed fair. That is, meanings that are symbolically violent negate our experience and feelings. And thirdly, the authority appears to come from a source other than its true one. From God or some moral or professional source, rather than being delegated from less visible power structures of church, caste or class (Wilson, 1991, p26).

These are specific examples of Bourdieu's (1990a) more general proposition that

> Every power to exert symbolic violence, ie. every power which manages to impose meanings and to impose them as legitimate by concealing the power relations which are the basis of its force, adds its own specifically symbolic force to those power relations. . . . . All pedagogic action is, objectively, symbolic violence insofar as it is the imposition of a cultural arbitrary by an arbitrary power (p4,5).

Bourdieu shows that pedagogic action reproduces the dominant culture in two senses; firstly because the power structure within which the learning takes place tends to mirror and legitimate, and thus reproduce, that of the dominant culture; secondly because the meanings inculcated have been selected (with corresponding exclusions) to reproduce the meanings of dominant societal groups. Both structure and meanings are arbitrary insofar as the structure and functions of that culture cannot be deduced from any universal principle, not being linked by any sort of internal relation to "the nature of things" or any "human nature"(Bourdieu, 1990a, p8):

> The sociological theory of pedagogic action distinguishes between the arbitrariness of the imposition and the arbitrariness of the content

> imposed, only so as to bring out the sociological implications of the relationship between two logical fictions, namely a pure power relationship as the objective truth of the imposition and a totally arbitrary culture as the objective truth of the meanings imposed. (p9) . . . authority plays a part in all pedagogy, even when the most universal meanings (science or technology) are to be inculcated. There is no power relation, however mechanical or ruthless which does not additionally exert a symbolic effect (Bourdieu, 1990a, p10).

Habitus

When a person has "lived" long enough through a period of inculcation of training, there is a durable product internalised by them which Bourdieu calls a habitus. Durable because it remains after the training has ceased, and is capable of perpetuating in practice the principles learnt. In this way the habitus produces and reproduces "the intellectual and moral integration of the group or class on whose belief it is carried out "(Bourdieu, 1990a, p35).

The habitus is a system of schemes of thought, perception, appreciation and action, a predisposition to "a rule-bound activity which, without being the product to obedience to rules, obeys certain regularities" (Bourdieu, 1990a, p64). Bourdieu (1990b) uses the analogy of the game to explain how the habitus functions:

> The habitus as the feel for the game is the social game embodied and turned into a second nature. Nothing is simultaneously freer and more constrained that the action of the good player. He quite naturally materializes at just the place the ball is about to fall, as if the ball were in command of him - but by that very fact, he is in command of the ball. The habitus, as society written into the body, into the biological individual, enable the infinite number of acts of the game - written into the game as possibilities and objective demands - to be produced; the constraints and demands of the game, although they are not restricted to a code of rules, impose themselves on those people - and those people alone - who, because they have a feel for the game, a feel, that is, for the immanent necessity for the game, are prepared to perceive them and carry them out (p63).

So the rules of the game construct the players, who in turn construct their own particular version of the game. And those who play the game the best are the winners who continually reproduce the game in its infinite variety, and create the illusion of freedom whilst the rules become ever more fixed, for

> The pedagogic work which produces the habitus . . . produces misrecognition of the limitations implied by this system, so that the efficacy of the ethical and logical programming it produces is enhanced by misrecognition of the inherent limits of this programming . . . The agents produced by pedagogic work would not be so totally the prisoners of the limitations which the cultural arbitrary imposes on their thought and practice, were it not that, contained within these limits by the self-discipline and self-censorship (the more unconscious to the extent that their principles have been internalized) they live out their thought and practice in the illusion of freedom and universality (Bourdieu, 1990a, p40).

Bourdieu (1990a) here demonstrates how difficult is to question the principles of one's own culture, for the very questions have their roots in that culture (p37).

Summary - power relations and standards

In this chapter I have started to reveal the backdrop for our drama, those social and political fields in which the human actors are enmeshed. The focus was on power relations, and the way in which they both violate and produce those who act out their lives within their pervasive influence.

In particular the mechanism of disciplinary power relations was examined, and the part that the normalising gaze of the examination has in controlling the players, and creating the modern individual as its supreme production; an individual defined by a competitive profile, an object positioned, classified, and articulated along a limited set of linear dimensions.

In the next chapter I show that crucial to this extremely efficient mechanism for achieving social stability is the scalpel that defines the classification that produces the person that lives in the house that disciplinary power built. A scalpel labelled standard!

Return to Table of Contents

Chapter 5: Power relations in educational systems

Synopsis

In this chapter, I take the more general ideas about power relations discussed in Chapter 3 and apply them to educational systems and institutions; in particular I unearth the many small social control mechanisms that pervade the school, and what sorts of people are produced by those mechanisms. I then examine the examination; how it normalises and individualises, and how it is impotent without the notion of the standard, the sword that excludes and rewards, the wedge that produces the gaps.

That brings us to the focus of this thesis, the suppression of error. There is a field of educational scholarship devoted to educational evaluation and measurement. Thousands of books. Hundreds of Journals. Most of the literature in the field is about errors in measurement. And of course, errors in measurement imply errors in the measurement of standards. Yet in classrooms and universities and public examining boards, on school reports and graduation and proficiency certificates, there is a great silence. It is as though this literature did not exist. Even prestigious testing agencies skim the surface of the error issue. The question is why? Why this suppression of the obvious empirical fact that educational standards as a thin accurate line have no empirical existence? It is to this question that the remainder of the chapter is addressed.

I examine the crucial part that the standard plays in the whole mechanism of defining cut-offs for abnormality and non-acceptance, and how important it is that these standards be seen as accurate if current societal structures are to be maintained.

Restrictions, penalties, productions

In the day to day operation of the school the power relations are activated through an array of petty restrictions and micro penalties, unrelated to the supposed primary function of the school as an institution designed to maximise learning. In most classrooms the policing of these restrictions takes a considerable amount of teacher time and often consumes more physical and emotional energy than does their teaching function. In many large High Schools in Australia, the major activity of the Deputy Principal is to deal with children with whom teachers are having disciplinary problems. We are obviously dealing here with what is a major part of the school curriculum, regardless of whether it appears in the official statement of syllabus.

There are restrictions on appearance and dress; on what may be worn, and how long or short it is; whether this be skirt, shirt, pants, hair, necklace, ear rings - whatever differentiates from the norm; whatever distinguishes an idiosyncratic persona; whatever, by whatever means, makes a public statement about personal autonomy. The restrictions will not be specified in detail, for fashions change too fast for that, and student creativity is limitless. However, the judgment of the school is, in retrospect and by definition, impeccable in these matters, and their verdict will rarely be contradicted, and never successfully challenged, by students. (or parents, for that matter). Significantly, school spirit, cooperation, health and safety, economy, equality, fraternity, are all likely to be part of the supporting ideology. But never conformity, for this would contradict the school ideological aims of developing individuality and autonomy. Yet surely conformity is what is being produced here; conformity, and the acceptance of the social sanctions that non-conformity bring.

Body, movement, speech and relations must be decorous: body and clothes must be not only clean, but tidy. Movement is both restricted and restrained: students should remain seated and never run in the corridors. Speech should be proper: slow, well-articulated, free of slang, swearing and salacity, respectful in address and tone, and preferably in the dialect of the upper middle class. And social relations should be moderate, free of all excesses; of love or hate, of enthusiasm or alienation, of spontaneity or cliquishness, of autonomy or dependency.

As well as physical and emotional containment, there is temporal curtailment. Work is restricted to what the timetable dictates. Maths must not be done in the history lesson, history must begin at 10 am., and no one may visit the toilet until 12.50 pm, unless they shame themselves by asking permission, and then only maybe.

There are a whole range of penalties utilised to reassert the power structure should any of the multitudinous restrictions of the school be breached: further physical containment during recesses, deprivations of various sorts, petty humiliations such as standing in corridors or outside offices, threats and harassments of various kinds, and finally physical punishment, suspension or expulsion. In 1997 in Australia the most popular fashionable sanction is called "time out", a broad notion that contains various shades of physical isolation, and which schools insist is not a punishment. The penalties are really of no significance. It is the acceptance of the penalty, which reinstates the integrity of the power structure, that is important. It is important that some students rebel, so that the power relations might be demonstrated (Wilson, 1990).

So what is produced through these restrictions and penalties? What is learnt? First, temporal regularity. There is a time to start and a time to finish, a time to sit and a time to stand. And these times are planned and arranged and policed by others. What is learnt is that time is determined not by the imperatives of life as they manifest themselves, nor by any plan that might make for some personal production, but by the dictates of people in authority, by the demands of an institution.

Second, physical containment. There is a space to be and a space to sit, and sit, and sit. What is learnt is that the demands of the body are not important, and it is preferable to forget that you have one.

Third, emotional contraction. What is learnt is that the exuberant emotional and psychic field must be reduced to the physical limits of the body, so that feelings and emotions are pacified, and the self reduced to placidity.

And finally, what is learnt is that all this has nothing to do with the maintenance of power relations, or the production of a social being, but is an unfortunate addendum to another far more important purpose; a necessary prerequisite for effective learning of the knowledge specified in the school curriculum. What is learnt is to misrecognise the social function of schooling.

Illich (1971) summarises the situation, calls it for what it is, and sees only one solution:

> School prepares for the alienated institutionalization of life by teaching the need to be taught. Once this lesson is learned, people lose their incentive to grow in independence; they no longer find relatedness attractive, and close themselves off to the surprises which life offers when it is not predetermined by institutional definition. And school directly or indirectly employs a major part of the population. School either keeps people for life or makes sure that they will fit into some institution. . . De-schooling is, therefore, at the root of any movement for human liberation (p47).

The examination

Before accepting or rejecting Illich's ultimate solution, let's look more closely at some of the specific mechanisms that produce this "alienated institutionalization of life."

First we look more closely at the examination, and at the particulars of its function. Foucault (1992) certainly affords it pride of place among the mechanisms of disciplinary power which he elucidates:

> The examination combines the techniques of an observing hierarchy and those of a normalizing judgment. It is a normalizing gaze, a surveillance that makes it possible to qualify, to classify and to punish. It establishes over individuals a visibility through

> which one differentiates them and judges them. That is why, in all the mechanisms of discipline, the examination is highly ritualized. In it are combined the ceremony of power and the form of the experiment, the deployment of force and the establishment of truth. At the heart of the procedures of discipline, it manifests the subjection of those who are perceived as objects and the objectification of those whom are subjected. The superimposition of the power relations and knowledge relations assumes in the examination all its visible brilliance (p184).

The examination is the ceremony of ordering; it is the mechanism through which real people (and hence the world) is ordered, and held in order, in all of the meanings of that word. By doing this in a setting in which the person who establishes order is also the person who establishes truth through knowledge, the certainty of correctness is established, and the person becomes an object in the acceptance of their place in the line, in their acceptance of their uni-dimensionality, in their incorporation of their relative merit as an essential part of their beingness.

Of course the examination is also a crucial element in the construction of human cognition. It defines what are true and false facts, what is right and wrong thinking, and what are the acceptable limits of intuition and feeling. But we are more concerned here with social categorisation.

The report is the place where such individuality is made official; here is the permanent record, uncorrupted by any possibility of error, of one's place in the order of things; of a person's history, present, and future distilled into a single mark; of a sign that evokes possibilities and defines exclusions; in the world of higher education and the world of work, here is the official indicator of who you are, what you are.

Foucault (1992) indicates that this individualisation through comparison is intensified as power disperses and abnormality increases:

> as power becomes more anonymous and more functional, those on whom it is exercised tend to be more strongly individualized; it is exercised by surveillance rather than ceremonies, by observation rather than commemorative accounts, by comparative measures that have the 'norm' as reference rather than

> genealogies giving ancestors as points of reference; by 'gaps' rather than by deeds. In a system of discipline, the child is more individualized than the adult, the patient more than the healthy man, the madman and the delinquent more than the normal and non-delinquent (p193).

It is at these crucial points that define exclusion that any error becomes unacceptable. These are the points that define, not so much the norm, but the gaps that define abnormality, unacceptability, dangerous deviance. The normal is indeed defined by a broad grey band, but it is essential that the abnormal be determined by the thin red line that separates. And that line, that thin red line where the blood flows, is the standard.

Standards and swords

Foucault does clearly show how the battle lines are drawn up. He displays the deployment of troops and the strategy of the battle. With unerring accuracy he pinpoints the diversions and ambushes and the misinformation and propaganda that camouflage the major thrusts.

Even so, he pays almost no attention to the major weapon which ensures success, to the one notion without which the whole structure is unstable; he downplays the construction that turns a house of straw into a house of bricks, and allows that momentous separation between the good little three little pigs, and the big bad independent wolf. Could it be that his academic self wished to retain this last bastion of its own identity?

Regardless, without the steel edged standard to cut off the tail with a carving knife, and without the standard chippy chippy chopper on the big black block to lop off the heads that are too way out, disciplinary power is reduced to a shadow. The notion of the norm is dependent for its existence on the notion of the not-norm, on the notion of the abnormal. And the abnormal owes its existence to the act of separation.

Regardless of how disciplinary power is deployed, whether through the micro-penalties of day to day detail, or the graduation rituals of national examinations, or definitions of insanity, the thin line between the acceptable and unacceptable must be drawn. And it can only be drawn by evoking the idea of a standard, of an cut-off point that can be accurately determined and applied. All this regardless of whether we want to evoke democratic values, or scientific values, or aesthetic values, or other "expert" values in determining the standard, and then measuring it.

For without the notion of the standard there can be no classifications, no qualifications, no exclusions. There can be no norm, because there is no abnorm. There can be order, but without the standard there can be no disorder; Without the standard, we can still construct an order of merit, but cannot differentiate excellence, or determine exclusion; we can still individuate by placing on a line, but we cannot delineate winners because we cannot define losers. A race where everyone gets a prize is like a race where no one gets a prize; it loses its purpose as a race, and soon becomes a game that no one wants to play. Gilbert was right: "When everybody's somebody, then no one's anybody."

The blade must be sharp. There is no room for error. There is some aesthetic beauty, some notion of swift justice, black and violent as it might be, in a blade that cleanly and swiftly decapitates. Yet a mangled hatchet job will inevitably evoke horror. And so it is with any application of the standard. The acceptance of classifications and exclusions, both by those who apply them and those who are their recipients, are dependent on the precision and truth of the standard. Without these qualities the whole examination exercise becomes exposed as a political ploy to order and control, to reward and exclude, to hold in place vast structures of inequity. In short, it becomes exposed as a hatchet job.

A place to hide

If it is indeed true that the notion of standard is central to the maintenance of cultural identity as we live it, as central perhaps as was the notion of God to the cultural identity of life lived in the Twelfth century, then we must not be surprised that the notion is highly resistant to empirical contradiction. Nor should we be surprised that those who are aware of any such contradiction have some realisation of its traumatic nature, and of the necessity to keep it secret.

The human mind is remarkably efficient. Socially inclined as it is, it realises the only way to keep a secret is to hide it away. So the secret becomes a secret from one's own consciousness, locked away down there where angels fear to tread. The unconscious is nothing more than this; the space where we hide what we know from our conscious selves because the knowledge contains a truth that is too hot to handle, an awareness too destructive to life as we know it.

Would the social world we know really collapse if the notion of the standard had to go? Would we dissolve in chaos, or move gently onward to build a better world? Or would we simply find another subtly socially reconstructed lie to replace the one we'd lost?

Summing up

We have seen how central the notion of standard is to the maintenance of the social structures of power in which we are enmeshed, and to education's crucial social function of categorisation.

There are affect components involved here; the bearer of the standard is clothed in fancy emotional underwear, wears a colourful mythical costume, and carries a sceptre that denotes moral high ground. In the next chapter we examine some of these other dimensions of the assessment fairy tale.

Return to Table of Contents

Chapter 6: Standards, myth, and ideology

Preview

After a brief look at myths and rituals, and the special place they hold in our thinking - a place apart from critical thought, I assign the idea of the human standard as currently understood to this mythological sphere.

I look at the emotional intensity of discourse about the standard, its significance as an article of faith, a basic assumption, an ideological king-pin, and at who gains from the non-recognition of its problematic classification. Specifically, I show how the notion of a standard of behaviour in families helps to maintain the family structure; then I examine in some detail the mechanisms the school uses to maintain "emotional" standards by denying the reality of human feelings, and how this is related to the maintenance of control, of good order.

Flags

When the army begins to march, or the Governor returns to his residence, the event is heralded by the raising of the Standard. The flag is the symbol of their power. When we salute the flag, we do obeisance to that power, in which glory resides. And, when power is embedded in the relationships of human structures, we salute the standard, we pay homage to the strength of those structures, simply by our willingness to play our designated part within them; in short, by our subservience to structural dictates, and our acceptance of relational obligations.

This language is hard to live with, this description too intense for comfort. We need a softer cushion on which to fall, a more prophylactic myth to justify our allegiances and comfort our losses. As we shall see, we will find such justification in the world of moral values.

These relational structures often have no visual symbol to represent them, though particular versions of them proliferate in the form of corporation logos, school and family crests. These are usually of limited emotional impact. More successful have been brand names for clothes, where the image behind the symbol has been so successfully assimilated that not only are consumers willing to pay much more for the product, but are proud to become walking advertisements. Some Japanese corporations and some sports teams have managed to construct songs that fit the bill. But in general the "flag saluting" within families, schools and workplace has been accomplished more through particular discourses with words and body language than through responses to visual symbols.

Discourse and value myths

I use discourse here to describe not only "what can be said and thought, but also about who can speak, when, and with what authority. Discourses embody meaning and social relationships, they constitute both subjectivity and power relations"(Ball, 1990, p2). Discourses thus constrain the possibilities of thought, and are defined by what is absent from them as much as by what is produced through them.

So what are the key elements of discourse around standards? What are the words and phrases that trigger a "flag" like response? For whilst it is true that most social structures can, if necessary, muster some physical force - in the form of army, police, courts, psychiatric hospitals, masculine muscle - to deal with minor perpertations of the structure, the inherent strength of the structure is vastly greater than such disciplinary mechanisms that may be utilised. Just as in a crystal it is the individual molecular bonds which bind the crystal in its hard, rigid and determinable form, so it is the acceptance and actioning by each person of the appropriate relational roles between people that account for the maintenance and solidity of the social structure. So how constitute a symbolic reminder, a conditioning stimulus, a ritualistic nudge and wink, that stimulates and fortifies the memories of our proper relationships to those who lead us or are led by us, to those who love us or whom we should love, to those to whom dues are owed, or to whom we owe our dues?

The gross but honest dictates of parent-child relations are not effective with adults, or for most children for that matter, raising as they do so much overt rebellious reaction. "Do what you're bloody well told" does not trigger the appropriate response. The linguistic flag carries much more powerful symbols in its armoury. Looking upward, we see Duty, Loyalty, Respect, Discipline and Strong Leadership all emblazoned on the High Standard in gold letters. And looking downward, the cold sharp chisel of Efficiency nestles neatly in the caring hand of Institutional Love.

It is important to understand that once these abstractions are incorporated into a personal value system, so that they become part of a way of being, a way of institutional living, the ground of faith on which hierarchical life is premised, then dependence and obedience all become responses that inhabit moral high ground, for they are necessary to maintain, not the hierarchy, but the values in which it is now delicately clothed. And the violations they entail work efficiently underground in this hallowed space.

Further to this, the more intense and horrible the violations involved, the more pervasive and enduring the myths and values that provide the cover up and justify the carnage. The Freudian myth embodied in psychoanalysis regarding the sexual fantasies of children is a good example. The myth enabled child sexual abuse and incest to be disguised and trivialised for a hundred years, as we are only now beginning to realise; sexual abuse of the child became translated through therapeutic discourse to sexual fantasies of the child aimed at the adult (Masson, 1991; Miller, 1984). The myth of the glory of war has required the joint barrage of visual human slaughter on television, together with an appreciation of the probability of global nuclear extinction, to diminish its insidious hold on our thinking. And even now the monster will not lay down and die.

And there is another aspect of enduring myths that we must not forget. Such myths do truthfully represent a part of the human condition. Many children do sometimes act seductively towards their parents. There is a form of transcendence in the self sacrifice and comradeship that is a part of some men's experience of war. Yet when these myths are used to disguise the carnage, rape and pillage that are their major manifestations, then such myths become not the harbingers of truth, but their disguises.

What I am asserting in this thesis is that the myth of the human "standard" is just such a myth in the more "civilised" wars of structural violation in which our lives are embedded, wars no less destructive of human life and potential because their weapons are so insidious and subtle: Wars to which at this time in our history it is now appropriate to turn our attention, so that we may, in a non-violent way, bring about their cessation.

Standards and discipline

Talk about raising educational standards evokes intimations of glory and solidarity, of battles won and lost, of remembrance of our dependence on elite leaders and arcane specialists. Who talks of the shocking implications of lowered standards and the necessity to keep them high, and to whom do they talk? Who are the flag-bearers to defend us from the horrors of mediocrity, and the hellish consequences of the (inevitable) average? What do such utterances herald, and do what do they respond? (Wood, 1987, p214).

In the public arena, whether that be the political castle of public affairs, the media circus of public relations, the disciplinary field of the public service, or the common ground of the public house, talk of raising standards is invariably linked with the idea of better discipline. Contrarily, the cause of lowering standards is clearly tied in public discourse to soft leaders and the inevitable anarchy which that is fantasised to produce.

So "standards are also values to which people aspire or lament the decline in or lack thereof." (Norris, 1991, p335). People talk about raising standards when they perceive a slackness in the ropes of control, when they see a sloppiness infiltrating the verities of life, when they begin to be fearful about life's diminishing certainties. Talk of standards is talk about conservation, about protecting the past in its imagined superiority and security, and defending the future through strong leadership. "Discipline," "Respect," "Standards," "Leadership" are almost interchangeable words in a discourse that lauds the good old days and decries the soft underbellied freedom and license of the present. It is the language of the old talking about the young, of the powerful talking about the rest of the world, of the mind talking about the body, of men talking about women. And these days, let us be fair, of some women talking about men. By implication, it is discourse that defends appropriation and privilege, and the structures of inequity in which they flourish.

Suffering together

Heraldic and educational standards both also share a deep emotional component, digging deeply into the well of group identity that tribes and political parties, multinationals and nation states, know so well how to bring bubbling and boiling to the surface. We all know the clarion cries that activate the emotional unity that is evoked and manipulated by demagogues - the Fatherland, the Motherland, Our Land, Our Nation, Our Church, Our Family, Our Team, Our God, whatever its particular form. Words that recall our common heritage and our common destiny, and the myths and ideologies that surround that communality; we lose our individual and insignificant identity in the power and communion of the group, and are seduced into forgetting our fear even as we lose our freedom.

Through such languaging the notion of standards and their conservation becomes emotionally tied to our deep sense of wanting to belong, wanting to have our place in the social world. And of course, our place in the social world is dependent on the survival of that social world in which we have our place.

At the very least, discourse about standards will be emotionally charged. Talk of changing educational standards is like talk of changing the flag. It triggers all the fears of change in the social realities, be they ever so violating, for which the standard, and the flag, are symbols.

By insisting in this thesis that educational or ability standards have no empirical reality, I cut much more deeply into the social fabric. For such a claim not only undermines the standard, but also by association denigrates the social reality that it represents. The metaphor is not changing the flag, but destroying it, on the grounds that the social order that it pretends to represent is a delusion, very different to the one that it does indeed refer to. A delusion whose continuance, furthermore, is largely sustained through the emotional effects of the inviolability of its recurring symbol, the flag.

The person who destroys the flag is inviting extreme social response, for such is its emotional content that many people will identify this map with its territory. For them, to destroy the flag is to destroy the social order it represents, and thus to destroy their identity within that order. Emotionally, social symbol and social reality are contiguous. For many people, this contiguity overlaps and symbol and referent become identical. In this state of mind, cognitive arguments and empirical data have as much impact as falling animals crashing into rocks. As much impact on the rocks, that is.

In an analogous way, to criticise the notion of educational or job standards on the grounds that they cannot in practice be measured or logically sustained is to destabilise the symbol of the meritocritous society, the competitive capitalist order that it supports, and the cult of individualism that, almost alone, it defines and constructs. Emotionally, these four constructs - standard, competition, meritocracy, and individualism, are deeply intertwined. To threaten one of them is to threaten all. And to threaten all is to threaten each one of us, you and I and him and her. For it is to threaten that social order in which we all, in our own way, or more likely in a way that the structure has imposed on us, has found our place.

Fact or faith - the sociological imperative

So the standard is a social construct whose meaning is not dependent on any empirical evidence to support it. The flag is not a bit of cloth attached to a pole; it is an idea, a social construct, with which most of us, individually and in a group, interact in fairly well-defined ways. In a similar way, money is not a piece of paper with pictures and writing on it. It is again a social construct which most people are willing to agree has a certain meaning which includes an intense emotional component. But again, a social construct dependent on faith for its continuance. Lose that faith, and the value of the money evaporates.

Likewise the notion of a standard: It is a notion, an idea, a social construct that helps bind together the social structure that brings order to our lives. If, as I have suggested, it is a very fundamental construct, one which is central and crucial to other social constructs which in this time and place are thought to have particular value in constructing (and thus validating and justifying) the social relations in which our lives seem inextricably enmeshed, then even more reason for letting it alone, for not subjecting it to too critical inspection, for not undermining a fundamental article of faith.

Articles of faith do not need empirical evidence to support them, and are extremely resistant to empirical evidence that casts doubt on their logical consistency or their stability or their contradictions to other articles of faith. For articles of faith tend to develop around themselves other ideas and ways of relating that are reasonably consistent with them. These coordinations then constitute a way of living in the world, a set of habits that helps give a sense of stability and thus timelessness in a world in which change is inevitable on every street, and chaos is just around the corner. They constitute, in other words, what we call social reality. They might more accurately be called the social fantasies we construct and live that help make the conditions of our lives, and the lives of selected others, more bearable.

And if this cuddly teddy bear turns out to be a real dragon, destroying the lives of many more than it supports, then all the harder to slay it.

The psychological imperative

When we are dealing with the educational assessment of students we must add the teacher's psychological necessity for accuracy. At some level teachers all know how important their assessments are to the futures of their students. They all are aware of its use in social stratification, and its more negative function of the excluder, and the destroyer of personal dreams. And this mechanism operates through self exclusion as much as exclusion by any external force.

This is the load the assessor carries: for the students themselves usually accept the judgments made of them, and compose their lives accordingly. This is self imposed as much as it is dictated by any external agency. So through their assessments, teachers have monstrous effects on the future lives of their students. This is an acceptable load if the assessments are very accurate, and do in fact measure the capability of the student. But if they are enormously in error, what then? What is the psychological price of instigating massive inequity, enormous misplacement?

Instrumental value

The notion of "standard" has a particular function in the value conglomerate of respect-discipline-efficiency that is a major part of the ideological glue that helps hold hierarchical systems firm. For the standard is the value that mediates between ideology and structure, between the moral values, and the relational power systems that they support. The standard defines the point of action at which any disjunction between value and experience is challengeable.

Let's see this in action in two hierarchies; first in relation to respect in the home; then in relation to emotion in the school.

The family

In a family, duty, obedience, respect, discipline are continuous, rather than binary, constructs. That is, children are more or less dutiful, or obedient, or respectful. One child is more disciplined than another. So how do we know when we reach the point where acceptability is breached, where unacceptability is reached? We know because what has occurred is below the standard. As parents we "know" there are standards of behaviour that must be observed. And the disciplined child is one who knows, accepts, and behaves within the limits of these acceptable standards. And these standards are not of my making as a parent, but something that "society" demands. I may have very high standards, in which case I may be tougher (and hence more moral) than most others. Or I may be softer (and hence more humane or emotional) than most others. But the myth of a "standard", that point of demarcation between acceptable and unacceptable, is implicit in both these positions. And my duty, as a parent, is to maintain this standard.

That this standard has no empirical stability (certainly not for the group and generally not for the individual) is insignificant in the light of its logical necessity to maintain the structural stability of the family. After all, how can a parent ever demonstrate the extent of power difference if that difference is never confronted with an explicit, implicit, or fantasised challenge?

Sexuality and school

The hierarchy that is the school is much bigger and less personalised, so is harder to hold firm. So there are many standards of behaviour to hold emotion in check, and many standards of cognition with which to gain leverage on the mental processes. This is equally true for both teacher and student. We like to make an ideological separation between school discipline and the school disciplines, yet the processes by which each are engendered are similar if not identical.

So how are emotions in a school controlled through the imposition (or better still the personal incorporation) of standards? Firstly there is the professional standard of distance, of objectivity, of detachment. Emotional involvement, whether positive or negative, is taboo. Professionally the emotions are controlled by pretending that they do not exist. On the positive side the standard is that low level of affect described as "friendly interest." For young children this may be expanded to "fondness" unless you are male and the student is female. On the negative side the standard, the limit of negativity, is a low key sternness that accompanies correction. Essentially these low level affects are seen as acceptable nuances of cognitive behaviour.

Neither anger nor love have any place within the professional role of the teacher. To indulge either is seen as a breach of professional ethics. Such standards are justified by claiming that any relationship with students involving emotion would be dangerous to the students involved and unfair to the others. Dangerous because escalation could lead either to violent or sexual outcomes. An example of the catastrophic consequence justification. This disguises the stronger and more immediate danger, of course, which is to the stability of the power relations. Legitimate anger at the inequities hidden in that structure, or of love that transcends it, both pose fundamental threats to its continuance.

For the student in school emotions are also ignored. They have no place and so do not exist. Any acting out of emotions however is given high priority and the school disciplinary structures are immediately brought into play. The emotions are ignored, but the behaviour is punished. This is equally true regardless of whether positive or negative emotions have inspired the behaviour. Indeed, the school authority is much more comfortable with handling the acting out of negative feelings of fear or anger or revenge or envy than it is with any overt expressions of love or sharing or student cohesion, so easily interpreted as solidarity and hence politically suspect as potentially destabilising.

Emotional intimacy between students, or between a student and teacher, is rightly seen to be incompatible with the power relations that define the school structure. Two students who actively demonstrate their passion are likely to be dealt with more harshly (probably by expulsion) than are those who actively act out their hostility. Hostile students allow the school to demonstrate its own power. Loving students can only highlight the emotional vacuum of the school's structure; and incidentally expose the obsession with sexuality that underlies its prohibition. That the taboo is so seldom breached is evidence of the school's enormous power, especially so during adolescence, where for many students it is their major preoccupation.

Demonstrated or inadequately disguised love between a student and teacher, even if completely non-sexual in its overt manifestation, evokes a response amongst teachers almost as powerful as the response to incest. Outside the context of the school, love between people of different ages is an accepted norm, so long as the differential is not too great. Within the school context, it is condemned on the grounds that it is an abuse of power. The assumption is that the teacher has abused his or her power over the student and manipulated the student's affection. Now whilst this may be true in some circumstances, and whilst the roles in the school have doubtless influenced the relationship, intense emotional relationships that develop between the two people (rather than between their partial selves in role) are much more than this. They are as common and as intense and as potentially fulfilling as are such relations occurring in any other social context.

To understand the strength of the taboo we must understand that it is not so much the abuse of power that is involved here, but its elimination, its disintegration, its transcendence. Love and power are incompatible relations (Laing, 1967). Love is a state of openness and mutuality in which the other is accepted in his or her wholeness, where there is trust in the flow of positive affect, of cohesiveness. Control is the denial of such trust, and structures defined by hierarchical power relations are thus structures permeated by mistrust (Maturana, 1980). Hence the necessity to control and punish.

So love relations between a student and teacher are not taboo because they might lead to sexual relations, or because they are unfair to other students, or because they represent an abuse of teacher power, or even because they might represent a malicious manipulation of the teacher by the student. Or because of the many additional justifications for the taboo that we could construct and fantasise. All would possibly at times contain some grain of truth, and all would miss the target by rendering it invisible. The fundamental immorality of such relations is that they are contradictory to the structure of the school, to its defining power relations, and are thus a fundamental threat to its continued existence.

It is equally important to understand that this fundamental reason for the taboo will be disguised in any particular case by evoking the concept of standards. The teacher is at fault because she has breached a professional standard of conduct which involves the abuse of power. The student will be at fault because he has not realised his vulnerability and has not allowed himself to be sufficiently protected by the benevolent authority which has defined the standards of student behaviour. Like so many rules in a school, this one, about loving teachers, does not appear in the rule book. Even so, no student would truthfully claim they did not know that it breached the standard of acceptable behaviour. And few would be able to rationally justify its abolition.

As described earlier, the appearance of the standard invokes an emotional response rather than a cognitive one. It bypasses notions of equity or justice that might grow out of a rational debate on the power-control issue, on the limitation of personal freedoms. It sidesteps any possibility of an ethical discourse by asserting that a standard has been breached, and thus by implication some act at the best unsatisfactory, and at the worst grossly immoral, has occurred. As the interpreter of standards, the school authority no longer seems to punish in order to defend its unequable structure. It now punishes in order to defend a high moral principle encased within "society's" standards. A violation of human rights has become a defence of all those things that "society" holds sacred, which become classified under the general rubric of "responsibility." And the use of the "standard" is the primary mechanism through which this mystifying ideological scam is accomplished.

Mind games

So far I have been concerned with discipline, with the way the school deals with unacceptable behaviour. Yet in educational discourse this is considered an unfortunate by product of the school's function. School discipline is defended not so much in its own right, but merely as a prerequisite to the maintenance of the disciplines. After all, the "real" reason children are at school is to gain knowledge, to become adepts of the various disciplines. Such learning, it is claimed, is dependent on the production of order, so that any control function that the school has is there to maintain the order that makes learning possible. Children are punished in school not so much for their own sake, though "god knows they must learn to be responsible for their actions", but rather for the protection of others. All must accept the discipline so that all may learn the disciplines.

Taken as an assertion about the nature of human learning, this is ridiculous. To assert that the best way for children to learn is to sit them down at desks in a teacher dominated classroom containing thirty or forty other children and change to a different topic every forty minutes is to deny most of what we know about the variety of learning styles and efficient learning environments. It denies a hundred years of research about how people learn.

Yet still the statements about good order, which in practice means being obedient and conforming, are central to the school philosophy. The reason is that such claims are not amenable to educational discourse. They are political statements, not educational ones. They are ideological statements designed to preserve the structure, and not therefore touched by empirical data. As articles of faith, as fundamental assumptions, they are flag waving slogans, amenable perhaps to emotional manipulation, but not to rational discourse.

All of which is not to deny that in an authoritarian-dependency structure, good order is necessary for effective "syllabus" learning to take place. It is, of course. But beyond that, and more pervasively, it is that structure itself that is inimical to learning. And it is largely in reaction to that structure that disorder occurs.

The ideology of order is necessary to protect those power relations from the dangers of rational debate, and the destabilising effect of empirical information that such debate might make visible.

Teacher stress

This ability of the system to protect itself from destabilising influences is nowhere better demonstrated than in the matter of teacher stress.

While teachers "stress out" in droves trying to maintain order, this is considered a second order phenomena. Their "real" function is to teach knowledge and skill, and school authorities consider it unfortunate that personal deficiencies on the part of the teacher might cause them stress.

In South Australia, "Stress Leave" is only available to teachers who are classified as "sick". Stress is a deficiency label attached to the teacher, a medical condition divorced from relational life. It may not be claimed by describing either the overt or covert violations within the structure of schooling, or by explaining it as attributable to professional or personal conflict with managers or students. The price of obtaining stress leave is the absolving of the institution for any part in its causation. (Section 30: (2A), Workers Rehabilitation and Compensation Act, 1986, South Australia)

Standards and destabilisation

We have seen how the notion of standard is a crucial ideological and mythical element in the hallowed structure of society. And an essential characteristic of the standard for that purpose is that it can be accurately defined and measured. In fact, standards can sometimes be defined and measured, but the errors contained in such measures are very large. I will show that they are in fact much larger than the massive literature on educational measurement and evaluation suggests.

Regardless, the notion of error is intrinsic and fundamental to any notion of measurement, and hence to any notion of measuring a standard as it is understood in the academic literature. Singer (1959) goes so far as to claim that "while experimental science accepts no witnesses to matters of fact save measurements and enumerations, yet it will pronounce no verdict on their testimony unless the witnesses disagree" (p101). So experimental science requires differences in measurements before it can decide what the "best" estimate of the measurement is, and the very notion of measurement is predicated on the notion of error. On the other hand any error in measurement is unacceptable if the notion of standard is to fulfil it's societal function in the categorisation of people. Who would accept failure or exclusion on the basis of a mark of 49 percent - plus or minus 15? Or even plus or minus one?

The simple professional and ethical solution is to attach an estimate of error to every application of a measurement of the standard, a habit deeply ingrained into practice in the physical sciences. However, this so contradictory to structural stability in the social world that to my knowledge the issue has never been seriously raised in professional debate about examinations, and when on rare occasions "ability" scores are presented as bands rather than lines they are based on reliability rather than validity considerations, so are gross under-representations of error; they are fudged instrumental errors, rather than errors in assessment.

Summing up

The standard is a crucial part of the assessment myth that is central to the stabilisation of power structures in modern societies. As such, attacks on its integrity, the naming of the gross errors attendant on its measurement, and explications of the violations to individuals that accompany its use, will be resisted.

Notions of standard have a very high emotional charge, and those who defend standards inhabit the high moral ground, as they defend the faith.

So challenges will be rare, and will be seen by most people as immoral, because they threaten the social fabric.

In the remainder of this thesis, one such challenge will be mounted.

Return to Table of Contents

Part 3: Tools of analysis

Chapter 7: Four frames of reference
Chapter 8: Equity, frames and hierarchy
Chapter 9: Instrumentation
Chapter 10: Comparability
Chapter 11: Rank orders and standards
Chapter 12: An inquiry into quality

Chapter 7: Four frames of reference

Synopsis

In this chapter four different frames of reference are defined; four different and largely incompatible sets of assumptions that underlie educational assessment processes as currently practised.

First is the Judges frame, recognised by its assumption of absolute truth, its hierarchical incorporation of infallibility; second is the General frame, embedded in the notion of error, and dedicated to the pursuit of the impossible, that holy grail of educational measurement, the true or universe score; third is the Specific frame, which assumes that all educational outcomes can be described in terms of specific overt behaviours with identifiable conditions of adequacy, and what can't be so described doesn't exist; fourth is the Responsive frame, in which the essential subjectivity of all assessment processes is recognised, as is their relatedness to context. Here assessment is a discourse dedicated to clarification, rather that the imposition of a judgment, or the affixation of a label.

Mythology

In the myth of meritocracy the examination is both a major ritual and a significant determinant of success. At the heart of this ritual, between the practice and the judgment, between the stress and the carthasis, is the great silence, the space where the judgment is processed.

The myth gives hints of what moves in this silence, for the myth makes three claims: the race is to the swiftest; the judgment is utterly accurate; and success is a certification of competency.

These hints tap the bases of the three frames of reference for assessment that assume objectivity. However, other assumptions of these frames make them mutually contradictory. This in itself would be good reason for keeping the process implicit. For the assumption

276

that inside the black box hidden in the silence is a mechanism, an instrument of great precision, may be difficult to sustain, if it contains major contradictions within its workings.

Four assessment systems, with four different frames of reference, have staked their claim to exclusive use of the black box, their claim to be the best foundation for the precision instrument to measure human - what? Bit hard to say what exactly. To measure, perhaps, human anything. It may be sufficient just to measure. Or even just to pretend to measure, to assert that a measurement has been made, so that a mark may be assigned to a person.

Frames, myths, and current practice

The Judge's frame is far more often evoked than talked about. The focus is on the assessor's judgment of the product. The major activity is in the mind of the assessor. Such terms as expert and connoisseur are essential to the construction of the accompanying myth. Faith is the requirement of all participants. It is explicit in discourses about teacher tests, public examinations, and tertiary assessment, and implicit in all human activities that involve the categorisation of people by assessors.

The General frame is the basis for educational measurement, for psychometrics. The focus is on the test itself, its content and the measurement it makes. Such terms as reliability and ability are essential to its mythological credibility. It purports to be objective science, and hence independent of faith. As such the world it relates to is static, so there is no essential activity. It is explicit in discourses about educational measurement, standardised tests, grades, norms; it is implicit in most discourses about standards and their definitions.

The Specific frame is about the whole assessment event, and is the basis for the literature that derived from the notion of specific behavioural objectives. The focus is on the student behaviour described within controlled events; in these events the context, task, and criteria for adequate performance are unambiguously pre-determined. Reality is observable in the phenomenological world; the essential activity is what the student does. This frame is explicit in discourses about objectives and outcomes; it is implicit, though rarely empirically present, in discourses about criteria, performance, competence and absolute standards.

The Responsive frame focuses on the assessor's response to the assessment product. Unlike the other frames it makes no claims to objectivity; as such its mythical tone is ephemeral, its status low. This frame is explicit in discourses about formative assessment, teacher feedback, qualitative assessment; it is implicit though hidden in the discourses within other frames, recognised by absences in logic and stressful silences in reflexive thought. Within the confines of communal safety such discourses are alluded to, skirted around, or at times discussed; on rare occasions such discourses emerge triumphantly as ideologies within discourse communities.

The Judge

Most assessment in education is carried out within the Judge's frame of reference. The chief characteristic is that one person assesses the quality of another person's performance, and this assessment is final. By definition the Judge's assessment is free of error, and therefore any check of the Judge's accuracy would represent a contradiction of his function. So such a check is not only unnecessary, it is immoral, in that it is an act likely to destabilise the whole assessment structure by calling into question its most hallowed assumption.

The Judge's assessment may be verbal and on-site, eschewing numeration and a special testing context. However, performance is usually assessed with tests and examinations, with merit graded in some way. It is assumed that adequacy or excellence in performance is described accurately by the Judge. For this to be true, it must also be assumed that the test measures what it purports to measure, and that the marking, whether by the Judge or his assistants, is reliable. Again, therefore, checks of validity, that the test measures what it purports to measure, or of reliability, that the test will give the same result if repeated, are not only unnecessary, but are unacceptable and demeaning.

Judges must stand firm on the absoluteness and infallibility of their judgments, for this is the essence of their power, the linchpin of their role, the irreducible minimum of their function.

Thus they are duty bound to recognise standards, to perceive with unerring eye that thinnest of lines that separates the good from the bad, the guilty from the innocent, the excellent from the mediocre, the pass from the fail.

Talk to them of normative curves or rank orders or percentiles, all of which imply relative standards, and they will hear you out, wish you well, and with scarcely disguised distain send you on your way. In their absolute world such matters are irrelevant. They know what the standard is, and therefore their job is simple. Simply to allocate students, or their work, to various positions above or below that standard.

Set hard in a rationalist world view, this is a black and white world, a fundamentalist cognitive universe. The assumptions deny the possibility of reality checks, so the collective fantasy easily becomes the perceived truth, as human minds and bodies contort themselves to deny their more immediate experience.

So let us see what that more immediate experience might tell us if another frame of reference is chosen.

The General

The second frame of reference is called the General frame. I used to call it the generalizability frame, but that word has been hijacked by psychometricians. The general has been privatised and corporatised by mathematicians. The bird has been tamed and lost its wings. The general has become severely contained in mathematical armour.

What I am calling the General frame of reference is blatantly egalitarian and inherently relativistic in its conception, but has become constricting, reductionist and inequitable in its mathematical application. In one form or another it has dominated the academic literature in educational assessment for over sixty years. Within this frame is contained most of the received wisdom from thousands of studies in educational measurement and evaluation.

Its two initial assumptions are shattering. One Judge is as good as another. And all Judges are inaccurate. God is dead!

Now as Little Jack Horner understood quite well, you can't just stick in your thumb and leave it there. If you stick in a thumb you've got to pull out a plum or no one will say you're a good boy. And the plum was the third assumption: There is a stable rank order of merit. So there is a true score.

And there is a stable standard. It's just that, sorry old chap, it's just that the jury does it better that the judge. Or perhaps it would be more accurate to say that we measurement experts, we psychometricians, can do it, with the jury's help, much more accurately than you can.

> *Judge You can, can you?*
>
> *General Yep.*
>
> *Judge Whose assumptions are you using?*
>
> *General Ours.*
>
> *Judge Whose definition of a true score?*
>
> *General Ours.*
>
> *Judge Whose definition of error?*
>
> *General Ours.*
>
> *Judge And whose definition of standard?*

General Ours.

Judge And you say I live in a fantasy world?

General That's what we say.

Judge I rest my case.

A bit unfair. But more that a grain of truth in all that. Even so, let's put a little more flesh on the skeleton of the General.

There is a true score: This notion has implications well beyond the psychometric. It is assumed that we are not measuring what a person can do, but rather a sample of what the person can do. If we could measure all the things (exactly) then we could find the true score directly. But as we can't there will always be some random error. In other words, if we had selected a different set of tasks the person would have done, probably, a little better or a little worse. Or even (softly now) a lot better or a lot worse.

This is all pretty obvious when you think about it. In almost any area of human activity, or study, there are an infinite number of possible tasks that could be required, questions that could be asked, limited only by the imagination of the examiners. And obviously, in a test situation, only a few may be chosen, from which a generalisation can be made about the rest. But the more tasks chosen, and the more they are a random sample of the total possible universe of questions, the closer you can get to the "true score". Further, *your* choice is a biased choice. Different people will choose different samples with different biases. So again, the more people involved in the setting of the examinable tasks, the closer we get to the replicable rank order, and hence to the true score.

We can't just stop at the questions, however; different markers rate answers differently. So markers also have to be sampled.

And contexts affect the result. Physical setting often affects performance. Some will perform better at home, some at school, some in an unknown environment. Some produce better work when isolated, as in a "normal" test situation. Others require stimulation in a group, which approximate more "normal" work situations.

The interactional media is sometimes crucial. Some express themselves better with the written word; others are much more comfortable with visual, aural-oral or more physical communication. Meanings can be communicated through many sensory modes. So if we are concerned to assess understanding of some area we would logically need to check across all of these modes.

And the time is important. They might do it well before lunch, badly after; successfully today, unsuccessfully in a month's time.

So assessments are required (marks or grades or rank orders), in all these different ways if we are to get a true estimate of a person's attainment or ability.

> *Whoops*
>
> *Whadaya mean, whoops?*
>
> *I saw that*
>
> *Saw what?*
>
> *Saw you pull that card out of your sleeve.*
>
> *What card?*
>
> *That one with the word "ability" on it.*
>
> *I didn't pull it out of anywhere. I materialised it. I created it.*
>
> *You made it up.*
>
> *I created a useful concept. We all do it all the time.*
>
> *Useful to who?*
>
> *Useful to me.*
>
> *Why is it useful to you to make up a concept called ability.*
>
> *Because I've created a mess. A conglomerate of numbers based on myriads of interactional and contextual incidents. And I know how to turn it into one fairly stable number. But then I've got to write it on a label and pin it on someone.*
>
> *Why?*
>
> *Why?*
>
> *Yes, why?*

Well, if I can't pin it on someone then I would have done all that work for nothing, because it's obvious that although all these scores and grades were supposed to be measuring the same thing, they were actually measuring different things.

And you've got to have them measuring the same thing?

Obviously, otherwise I can't add up all the marks to get one stable mark, can I?

I suppose not.

So I made up a name.

Ability?

Ability.

And no doubt you specified the ability as being identical to the task area you were assessing?

Of course.

So ability is what the total (average) number is measuring?

Absolutely.

Relatively, you mean.

Yes, it would be fairer to say relatively.

And if you know their ability you know what particular things they can do?

No, I wouldn't say that.

Perhaps you know what particular things they can do better than someone else?

No, not that either.

What do you know then?

Well, if you were to take all the possible things that a person might be required to do in a particular area of activity that is more or less described by the ability, then you could say that, on average, and very consistently, a person with a high score on that ability would do better than a person with a low score.

Whoops, you've done another shift. All this information isn't *about the person. It's about the interaction of the person with the task with the assessors. How are you justified in pinning it on the person doing the tasks? Why isn't this information about the whole contextual community?*

Initially it is. But when we average out all the individual scores, they stabilise for each person. Regardless of the context, and regardless of the particular assessors. And the only other stable objects in the whole shebang are the people being tested, and the thing we're supposed to be measuring. So it makes sense. Ability is the stable label.

What does that ability score tell you about specific things that they can do?

In terms of specific tasks I would have to admit, if pressured to do so, that I could, from their ability score, predict very little.

So you began with lots of information about differences.

Indeed.

And you finished up with one bit of information and a name attached to a person. One bit of information about a constancy.

True.

You made a choice. You could have said that a student's true ability was all that variety of things that were very uneven and unstable and changeable. You could have said that the true description of ability was the collection, rather than the summary or summation, of all the information.

I could have done that.

And then the summary, the average, would represent a huge simplification, a reductionist symbol, a monstrous error, rather than a true score?

That follows.

But you chose to define the average, the summary, the abstraction, as the true score, and everything else as error?

Indeed I did.

How do you justify that?

Because the average gives a stable score, and a stable rank

> *order, and this enables us to make a clear classification of the student.*
>
> *And that's important?*
>
> *It's crucial. You could say it was the aim of the whole exercise.*
>
> *I thought the aim of the exercise was to describe a student's learning.*
>
> *Would you think the best way to do that was with a number?*
>
> *No.*
>
> *Well, then!*

I have tried to give some of the flavour of the General frame of reference here. To indicate some of its assumptions, some of the things it can do, and some of the things that it can't do. And it is apparent that one of the things that it can't do is give specific information about exactly what tasks a person can or cannot adequately perform.

I have also, in the spirit of this frame, fudged a bit. For example, the scores are not stable; they are stabler after they are averaged than they were before. As are the rank orders. But stabler does not mean stable; more reliable does not mean reliable; more valid does not mean valid. More of this later.

I have also expanded the conceptualisation of this frame well beyond most of the theoretical expositions in the literature. Such logical expansion does not lead itself to elegant mathematical modelling, however, so the fudging of psychometricians has reduced, restricted and simplified these concepts to a shadow of their full power.

The Specific

The third frame of reference for assessment defines the world of specific behavioural objectives, or specific learning outcomes, and, by implication if not practice, of the more fashionable criterion based assessment and competency standards.

Here we are far away from the religious world of the judge, and the pseudo-scientific world of generalised ability. Here is a technological space in which a spade is indeed a spade, and to Alice's delight, things are indeed what they say they are. Or so it would seem.

This frame of reference assumes that the task of assessment is to describe what can be done, under what conditions, and what constitutes adequacy. So there is only one correct description of performance, and that is the unambiguous learning outcome that is defined in advance. It is assumed that learning outcomes can be defined so clearly that there is no doubt whether a person has, or had not, matched behaviour to the outcome.

There is no problem here of matching objectives to curriculum, and curriculum to testing. The objectives are the curriculum are the learning outcomes are the test. A rose is a rose is a rose.

Here is the bright fluorescent material world of the technological fix. Reality defined as observable behaviour. A world where doubt and uncertainty is no more. A place of clear goals, purposeful activity, and attainable and unambiguous outcomes.

More than this. This is surely a political revolution. The power to certify or exclude is no longer in the hands of the omnipotent judge or the manipulative psychometrician. It is clearly with the student who can self-certify adequacy, and any intelligent bystander can check that the task has indeed been adequately accomplished.

The technique was first developed to train technicians quickly and efficiently during the second world war to do a limited number of very specific tasks, and follow through a finite number of carefully specified procedures. In this it was highly successful, and its overflow into the general training area, and the nebulous and vague syllabuses of education, was viewed with delight by many of those who wished a firmer base for guiding and assessing learning. That is, who wanted to control what people learn.

And it was possible to find in most areas of learning, in most specifications of jobs, in most definitions of curriculum, in most topics of study, some irreducible minimum, some particular aspects of performance such that we could say - well, if they cannot do at least these things to this level of skill, or if they do not know at least these particular facts, then we could never certify that they were adequate in this area of functioning. In other words, the frame proved to be very useful where there were a finite number of tasks that could be isolated and specified, with limits of adequacy defined.

However, there were two questions, one technical and one political, which shattered the image of specific behavioural objectives as a democratic panacea for education. The first question was - is it possible to specifically define outcomes in any area of interaction that includes cognitive or interactional areas involving any problem solving or analysis or synthesis. Any activity, that is, involving cognition of more complexity than low-level comprehension?

Note, however, that to ask this question is to step outside the frame. For the assumption of the frame is that all tasks are so specifiable.

And the political question - who defines the objectives? Why these particular tasks? Why this particular context? Of what significance this particular cut-off for adequacy? Have we solved the problems of reliability or adequacy, or merely hidden them behind a dense materialist behavioural smoke-screen, behind which shadowy judges, bureaucratically insidious, silently sit?

Again, to ask this question is to move outside this frame. Within the frame this question is not a contradiction, it is simply irrelevant.

The Responsive

The Responsive frame of reference for assessment is manifestly and covertly subjective: no longer are the descriptions and judgments attributed to the performance, the artefact, or the person. What the assessor says is no longer claimed to be a quality of the object produced, or the objectified subject that produced it. What the assessor says is claimed only to be what it indeed is - a response of the assessor to a particular situation or artefact; a verbalisation of a particular human response to an interaction; a construction of the person assessing that says certainly as much about the world view of the person assessing as it does about some abstract quality or behavioural skill of the object or person being assessed.

Within such a frame there is no question of a right judgment, of a correct classification, of a true score. The response might be sensitive or insensitive, sophisticated or ingenuous, informed or uninformed. The verbalisation of that response might be honest or manipulative, its fullness expressed or repressed, its clarity widened or obscured. It still belongs undeniably to the assessor, and the expectation is not towards a conformity of judgment, but a diversity of reaction. The lowest common factor of agreement is replaced by the highest common multiple of difference. The subject of assessment is no longer reduced to an object by the limiting reductionism of a single number, but is expanded by the hopefully helpful feedback of diverse and stimulating and expansive response.

As with the other frames of reference, this one rarely materialises in its pure form. In the evaluation literature it has gained some attention under the rubric of formative evaluation, which occurs during a course of study, a low status cousin of summative evaluation, the final judgment, that more macho space where the real battles are fought, and the important decisions are made. Even so, there is professional literature in plenty, and especially in the rhetoric of "teaching" rather than "assessment", that supports the idea of assessment as feedback and guide, rather than classification and judgment (Williams, 1967).

So it is in this diagnostic and formative function that responsive assessment has found its place; as part of the training program rather than as legitimate description of what has been learnt.

There is good logical reason for this. It is obvious that this frame is a direct contradiction to the Specific frame, in which there is only one description of performance required and that is defined in advance.

It is less obvious, but none the less true, that the frame contains, in its practical functioning, a contradiction of the Judge and General frames, for it denies implicitly the idea of the single accurate order of merit, and hence the notion of some true score, or of some inviolate standard.

There is a further contradiction built into the assumptions of the Responsive frame. For if, in attending to the feedback, the performance of the person assessed is indeed improved, then the quality of performance, the degree of skill, will be changed, and the "true score" will also be changed in the very functioning of the assessment process, making the accurate judgment immediately inaccurate.

It is important to the logic of the Judge, General and Specific frames that no learning takes place after the test, for otherwise the test result becomes invalid, and must surely be dispensed with. On the other hand, within the Responsive frame, it is expected that the responsive feedback from an assessor will interact with the performance and improve the quality of later work, at least in terms of that particular assessor.

In the Responsive frame, this is an act to be applauded; in the other frames, it is a worrying source of error; in this respect the Responsive frame fits into a dynamic, and hence educative, environment. The other frames are predicated on a static universe, and are thus, in a profound sense, anti-educational.

Shifting sands

How does the Judge perceive the other frames? To the Judge the General frame is hopelessly relativistic, lacking in authenticity and depth, and devoid of standards. the Specific frame is reductionist and trivial, unable to cope with the cognitive complexity which lies at the heart of any discipline. And the Responsive frame is permeated with that subjectivity that indicates the absence of the objectivity that only comes with true scholarship, which the Judge exemplifies.

How are the other frames viewed from the General perspective? The Judge simply cannot deliver his promise of measuring accurate standards. His idiosyncrasy is legion and his omnipotence is self delusion. The Specific frame presents information that is scattered, incapable of producing a single dimension of measurement. Any addition of the specific information loses it, and returns the data to the General frame without the usual measurement controls. The Responsive frame presents data that is too diverse and contradictory to be seriously considered as a measurement.

From the Specific frame the Judge may be measuring something but neither he nor anyone else knows what it is. Just so with the General frame, that gets lost in a wilderness of numbers and cognitive abstractions. And the Responsive frame belongs to the world of opinion and gossip rather than scientific description.

The Responsive assessor sees the Judge as a responsive assessor, deluded by a fantasy of objectivity and accuracy. The General frame is seen as mathematical chicanery used to justify unsustainable classifications of individual people. And the Specific frame is seen as an absurd attempt to reduce human experience and performance to a few describable and measurable behaviours.

Conclusion

Sensible debate within a particular frame of reference for assessment sometimes occurs. However, rational debate across the full range of frames is a rarity. Part of the reason for this is that people argue from different frames of reference, with their incompatible assumptions, and these are rarely made overt. Not only that, but individual people in a particular discussion shift from one frame of reference to another, sometimes with bewildering speed.

This is why a conversation between a university professor (Judge), a psychometrician (General), a educational software technologist (Specific), and a radical teacher (Responsive), sounds like the sound track from a Marx Brothers movie.

In the next chapter we shall see how these frames are related to concepts of equity and hierarchy.

Return to Table of Contents

Chapter 8: Equity, frames and hierarchy

Synopsis

In this section I want to tease out some of the relationships between equity and assessment.

Life wasn't meant to be easy. We have four frames for assessment. Four differing sets of assumptions about what assessment is about. Equity is similarly compounded. There are (at least) three differing definitions of equity in current use: The first is based on equal means, treating everyone the same; the second is based on equal ends, treating everybody differently to end up the same; and the third is based on elucidating different ends and different means. The advantages, limitations, and pre-conditions for these three notions to be effective in practice are discussed.

Then I take each frame of reference for assessment in turn, and tease out its compatibility with each notion of equity, and with the hierarchical power relations of which the assessment system is an integral part.

The meaning of equity

Equity means fair, says my dictionary. And fair means, you guessed it, equity. I asked my seven year old daughter what fair means. Sharing things, she said. Still not satisfied, I asked my five year old. Fair means not missing out, she said, being included.

That seemed like a good start. Notions of equal shares and inclusion. But the meaning gets more complicated as the implications for achieving fairness are developed.

Equal treatment

The soft definition of fairness is that everyone gets treated the same. But then they end up differently because different people respond differently to the same input. We can say that's fair because some people are more intelligent or work harder so we would expect them to gain more. But then if the nature of the input is changed, different people succeed. And the people who succeed often seem very similar to the people who design and implement the input. Not surprising really.

What has been designed here is a nice tight closed logical system; people design educational means and ends to produce people rather like themselves and also produce definitions of intelligence or ability or skill or relevant knowledge based on similar means and ends, thus justifying the fairness of the unequal ends in terms of the unequal intelligences of the people attaining them.

The self fulfilling prophecy continues when we make these unequal ends the criteria for selection to favoured occupations (Goslin, 1963 p156). Here the success of the incumbents, and all are deemed successful by definition once selected, proves the value and validity of the whole process. Certainly none of the people so favoured are likely to suggest that almost anyone could do their job given an appropriate training programme, or, even more unthinkable these days, through an informal apprenticeship.

How do teachers react to this soft definition of equity? For those who see their task primarily as transmitting certain knowledge and skill and attitudes to students the definition is appealing. Because they see their professional task as transmission, they are likely to define clarity of communication in terms of logic and intention rather than in terms of accessibility or effect. Thus their professional integrity will be preserved if they treat all students in exactly the same way. It will even be considered an advantage if all students dress the same in some sort of uniform so that personal idiosyncrasy is visually nullified.

At the other end of this spectrum are teachers, often those who teach very young children, who have some sense of the student as a person with a very particular background and learning style, and who have a sense of responsibility to deal with those differences, albeit with certain specified skills or knowledge as having particular importance. Such teachers will see the gross limitations of this equal treatment definition, and will tend to reject it.

Yet even these teachers are likely to be ambivalent about rejecting this definition entirely, because of their position in the total educational structure. After all, there is a curriculum that all students are expected to master, and the larger and more structured the organisational unit in which they are enmeshed, the more likely they are to feel the pressure and surveillance directed towards particular ends. And the bigger the group of students they are confronted with, the more helpless they are likely to feel about the possibility of treating everyone differently.

Then, confronted with the impossibility of treating the children differently, in confusion they abdicate: if it isn't possible to achieve equity of ends through differential treatment, isn't it best to at least achieve equity of means?

Equal ends

Let's take a closer look at this harder definition of fairness; fairness is treating everyone differently so they end up the same.

The reasoning is clear. People have different prior experience, so they necessarily start a new experience with different prior knowledge and skill. So if they are all treated the same, this differential starting point will produce disparate ends. It follows we must treat all of them differently if we are to give them all the same opportunity to reach the same specified end point. Fairness or equal opportunity thus means giving additional resources and time to those who are originally disadvantaged in order to achieve equality of ends.

Surely that's fair? Possibly. But who decides what these ends are that everyone should strive to reach? Usually they are defined by an unrepresentative group, who have a strong vested interest in maintaining and distributing certain sorts of knowledge, values, skills and myths, and/or of limiting the number of people who will have access to the same. Thus the ends are a narrow selection from a much wider range of possibilities. Why should all the resources go into these particular ends?

Part of the answer relates to the current nature of institutions, and the learning that can occur in them. They are not constructed or resourced in a way conducive to individualised learning, but in terms of much larger learning units.

So teaching institutions tend to ignore the unfair treatment of individual students for two reasons: First, because individual students have no power, this representation of unfairness is rarely articulated; and second, because an adequate differentiated response would administratively smell of disorder, such an approach would be contrary to the institution's structural purpose as a hierarchy, which is to impose order.

Some sub-groups however do have power. Institutions have to respond to claims of discrimination against particular sub-groups of gender, class, ethnicity, or whatever minority has found a voice. This has been useful in the short term as an awareness raising activity about the equity issue.

Such political activity on the part of sub-groups that have found themselves disadvantaged by current structures of teaching has resulted in some shift, at least in terms of rhetoric, towards the equal ends definition of equity. There has been some small acceptance of the idea that it is equity of ends rather than of means that should define equity.

However, the "equal ends" comparison has been applied to groups, not to individuals; the debate has been about whether as many girls as boys can join the power elite, and not about the individualised treatment that might allow all who so desire to be successful. As the debate is about the sharing of domination between groups, it largely ignores the domination within such groups. As such it is also about the sharing of violation, and not about its elimination.

Equal ends and the myth of the intelligent child

Action has been at two levels. One involves awareness raising, so that members of disadvantaged sub-groups are encouraged to attempt educational activities previously not sought; for example, girls to study mathematics or engineering.

The other action has been, not surprisingly, to attempt an economic fix. Just as economic health, on the current fashionable models, supposedly bears a long term relationship to standard of living and quality of life for all, so more resources for the "disadvantaged" sub-groups will supposedly produce more equitable ends educationally.

Such an approach ignores the relationship between means and ends. For if it is the means, in this case the particular form of educational environment, that has actually produced the different ends, then more of the same means is hardly likely to improve matters. Indeed, intensifying the same means may produce more discrimination. (Of one thing though we may be sure. More resources for the disadvantaged will certainly benefit those advantaged who have identified the problem, and have some solutions, preferably packaged.)

How could this be? How could an educational environment, created by professional teachers, produce negative results, increase disadvantage? Surely anyone with sufficient motivation and intelligence can succeed?

That's one myth that has always stood in the way of any real progress towards sharing and inclusion. Once you accept the idea of "bright" students and "dumb" students, and the notion that there is a direct causal relation between attitude and success, then inequities are merely a mirror of these individual variables. If girls don't do as well as boys it is either because they're not so bright, they're not motivated, or both. And poor kids are dumber than rich kids and that's why they don't do so well. It's obvious. It's genetic as much as anything. Rich kid's fathers are more intelligent otherwise they wouldn't be rich!

Teachers, armed with prejudicial expectations and judgments as well as assessment data, are often quite clear about who is bright, average, and not so bright in their class, a distinction not always so clear to the outside observer. I've talked to small groups of children in hundreds of schools. I'd often ask the Principal to select a small group of about twelve students, some bright, some slow (one of the in-words for stupid at the time). We'd sit in a circle on the floor in the library and talk about home and school and life and the future for an hour or so. At the end of that time I was never able to tell which of the students were supposed to be the "slow" ones. I suspected sometimes they included those who had made the most significant contributions, and the most profound comments.

The "blame the victim" ideology is pervasive in education, and is maintained through the closed logical system described earlier. Assessment procedures play a crucial role here. After all, the teacher is paid to teach. Yet the failure label is invariably attached to the student.

Different people, ends and means

Because both the common ends, and the means of attaining them, seem to contain within themselves the seeds of the inequalities we are trying to diminish, we can try a third definition of fairness.

Fairness is treating people differently so they can end up differently. And the different ends will be determined largely by the students themselves. Fairness than consists in providing different resources so that different people can achieve their own different end points, through their own appropriate means.

Is this individual choice and freedom not illusory? Surely expectations embedded in people's social class or gender will determine their choices, and so inequities of power and wealth will still be perpetuated?

This is not a light criticism, and the strength of such sub-cultural or individual expectations is great. However, this strength is diminished as the awareness and verbalisation of the imposed expectations increases. Sub-cultural expectations do not invalidate the logic of the "difference" definition. They do indicate some of the conditions for an implementation in accord with its purposes.

The professional rhetoric of education is concerned with ideas of "individual differences", of the "whole person," and of "clear thinking, rational man." Less so with the passionate, spontaneous, loving, emotional man, or woman. Even so, we might expect some professional support for the different ends and means definition. There is, however, an inherent contradiction between the structure of educational institutions and this idea of equity. So the learning reality rarely approaches the professional rhetoric.

The structure of the school is hierarchical and competitive. The revered qualities are conformity (called cooperation), emotional suppression (called rationality), and acceptance of absurdity (called maturity or respect). None of these qualities is necessary for effective learning. Indeed, all are inimical to learning beyond the trivial. Yet all are necessary for success in learning at a school, because the institutional structure, the political reality that pervades the learning institution, demands these prerequisite responses.

Such an emphasis on control and order is simply incompatible with the idea of young people (of any people) being the main determinants of what they learn and how they learn it. That would be seen by the institution as anarchy. And whilst some teachers would see it as professionally desirable, they would go on to add that "in reality, of course, . . . "

What they mean is that the imperatives of their professional ethic and of their hierarchical morality are different. And in such a situation the hierarchical imperative will hold precedence. Such political expediency is often mis-named "reality". It is more accurately called political obligation, the moral imperative embedded in the institutional power structure. When professional behaviour is not subservient to this obligation, any teacher risks exclusion from the structure. Professional survival is, in the unreal world of the institution, indeed dependent on political expediency.

Equity, frame and hierarchy

Four frames of reference for assessment have been defined; four professionally legitimate ways to describe educational performance, each containing different assumptions about the nature of the task. And each, no doubt, differentially appropriate for particular purposes. Professionally there is an obligation to attach appropriate frames to such particular purposes.

Then three definitions of fairness have been described; three morally justifiable ways to describe educational equity, each fraught with its own limitations, and containing its own implicit notions about the meaning of justice.

These notions of frames and equity come together and form a discourse within educational institutions which are almost invariably hierarchical in their power structures, and these educational systems themselves are embedded in wider societal structures of that very special form of hierarchy called bureaucracy. This is not the time and place to go into detail about differences between simple hierarchies and bureaucracies. At the risk of oversimplification, I will note here that simple hierarchies usually have an identifiable person, with describable characteristics, at the apex. Bureaucracies, on the other hand, are led by shadowy and replaceable functionaries. Personal idiosyncrasies in such functionaries are abhorred. One of their tasks is to await their inevitable replacement by robots with phlegm and aplomb (Arendt, 1969; Kavan, 1985).

Now I want to examine the compatibilities between these professional assessment options, meanings of fairness, and the social structure called hierarchy.

Hierarchy, equity and the Judge's frame

Assessment in the Judge's frame is quite compatible with institutional hierarchy. More than this, by fusing the professional and political aspects of function the assessment process both strengthens and justifies the structure.

Specifically, if the Judge is necessary in order that the student may be accurately assessed, then the hierarchical structure is necessary in order to achieve this educational requirement. In addition, if a Chief Judge is necessary to check, or at least ratify, the accuracy of Lesser Judges, then the next level of hierarchy, the Head of Department, is necessitated. And so on. Thus the illusion that hierarchy is necessary for educational purposes is maintained.

Because the Judge's purpose and power are both based on his or her claim to recognise the standard, the equal treatment definition of equity dovetails nicely with this frame. Indeed, the assessor's work is so much simpler if all students have been through the same educational programme, so all have had an equal opportunity to know or respond to the answers to the questions asked. Whilst Judges would deny the necessity for a rank order of students, they would all be willing to admit that their task is so much easier once the rank order has been produced. All they have to do then is locate the standard between two particular students, and the classification of all the other students automatically follows.

The equal ends definition of equity presents the Judge with no theoretical difficulties. In practice however there are great difficulties.

Whilst the Judges believe they can recognise standards, the research indicates clearly that they are capable only of assessing comparative performance, and the "standard" is inevitably linked to the sample of responses provided, as well as to some assumptions about the composition of that sample. For example, given a large sample with a complete range of student work, a Judge will assess some (or many) as being below the required standard. Later, given a sample containing only those assessed previously at above standard, the Judge will now assess some of these at below standard, especially if he or she assumes the sample is covering the full range (Hartog and Rhodes 1936).

So even if the equal ends definition were achieved with a given group, and through differential treatment they had all reached an adequate standard, according to some data, it is almost certain that the Judge will still assess some at below the required standard.

However, as explained earlier, equal ends doesn't really apply to individuals, but to sub-groups. It's the relative percentage of success between sub-groups that assumes importance for the equity watch dogs. In this regard Judges, being rational and aware beings, are often able to adequately attune their prejudices to the political requirements of their time.

If the equal ends definition of difficulty sets a difficult task for the Judge, then rationally the different ends and means definition presents an incomprehensible one. For how could one hundred completely different products, the outcomes of one hundred different curricula, be compared to a single standard? Surely only Judges of very high status, or extreme arrogance, would attempt such a task.

Faint heart made not fair Judge! To the Judge it's no harder than any other assessment task. The Judge is undeterred by the variety of products and purposes. The Judge's standard is inviolate. The Judge simply compares each work to this standard and the decision is clear.

However, to do this they must of necessity apply their own criteria for success, rather than that of the student. In so doing they would countermand the requirements of an educational program directed towards different ends and means equity, in which the purposes, and hence the appropriate criteria, and thus necessarily the acceptable "standards", vary from student to student. Luckily, such rational considerations rarely impose on the Judge's religious rituals.

Hierarchy, equity and the General frame

The General frame has found little acceptance within educational institutions. Despite the fact that most of the technical and academic literature of educational measurement refers to this frame, and professional testing agencies use this frame for both standardised tests and for grading students, its egalitarian overtones, at least in regard to assessors, has found little response within institutions, despite the overwhelming evidence that using this frame produces more stable rank order grading of students.

Let's look at this a little more closely. The General frame of reference assumes that any single examiner is prone not only to idiosyncratic error due to differences in criteria and "standards" with other assessors, but also to considerable reliability error in his or her own remarking. That is, they will give different marks or grades if they mark the same papers on different occasions, or if they mark different versions of the same paper at the same time. And not only that, but such errors are increased, not decreased, if prior knowledge of the student is available (generalizability errors, that is). And not only that, but that chief examiners are no better than any others in regard to such heinous errors.

All this would be bad enough, interfering as it does with the "right" of the teacher or lecturer to have ownership of their students, and to alone decide their future. But if the assessment input of any competent person is as good as anyone else's, then the whole hierarchical structure of the organisation is called into question.

Worse is to come. Some studies have found that groups of students assessing their own work are also able to get closer to the "true" score than are individual learned superiors. This is democracy run wild; this is destabilisation of hallowed structures; this is anarchy.

Of course, educational institutions can survive without their Judges, although the professional justifications evoked by their presence does wonders for institutional status. If Judges lose the Wars of the Gradings to professional test agencies, then so be it. There are still plenty of hierarchical tasks to be done in selecting syllabuses, administering tests, limiting admission, marking rolls, ejecting students, and so on.

Even so, removing the myth of the Judge from the ideology of the educational institution is pulling out its teeth, leaving it gumless in academia. The function of the school and university has always been equivocal. Rhetorically defined by its purpose of searching for truth and instilling freedom of thought, its practical purpose has been much more mundane - to conserve the culture by perpetuating its myths and reproducing its social and technical elements.

The risk with academics is that they sometimes take their rhetoric seriously, and actively try to bridge the gap between ideology and practise. Given the somewhat radical stance developed in some schools and universities in the sixties and early seventies, it is not altogether surprising that they should be milked of some of their power during the eighties and early nineties of this century. The economic cringe is obvious. But what more Machiavellian way of producing an academic cringe than by using their own research as justification for removing their Judges' power.

In regard to equity, the equal treatment definition implies some measure of competitive merit, and such a measure would certainly be "fairer", that is more stable and less dependent on the vagaries of particular assessors, if the General frame of assessment were used.

This frame would also be useful in relation to the equal ends definition if professionally normed and standardised tests were used as an end point for a satisfactory standard. However, it would be a mistake to believe that the test measured any pre-existing standard. Rather the standard is defined by a certain score on the test. The validity of any such measure is moot. And indeed, this very mootness has left a gap in which the Judge has been resuscitated. For who else is capable to legitimise an arbitrary cut-off? (See any Public examination manual).

The rank ordering procedures of the General frame are not appropriate to the different ends and means idea of equity, because the educational ends and means are individually negotiable, so there is no single "ability" or "trait" or "domain" on the basis of which the students can be ranked.

Hierarchy, equity and the Specific frame

The Specific frame is very compatible with hierarchy. It is the ultimate in accountability and order. Once the outcomes are defined, or the domain of study clearly enunciated, educational programs using computers can reduce the whole educational enterprise to central administrative control, thus bypassing the sometimes difficult professional and technical considerations that in the past have hampered managerial efficiency. New-style managers in particular, wanting clear outcomes and economic accountability, are likely to regard the Specific frame, into which the severely bastardised criterion referenced assessment and competency standards has been incorporated, as a panacea.

Advocates of this frame are likely to down-play, and underestimate, the differences between the equal treatments and equal ends definitions of equity. It's simply a matter of time, they say. Our objectives are clear, our programs are tested, and everyone can reach the desired standard if they try. Some are a little slower than others, that's all, so they will require a little more time. But, given sufficient time, everyone will succeed (Bloom, 1976).

This is facile. Different treatment involves much more than time. Learning styles and appropriate student-teacher relationships cannot be condensed into this single variable. None the less, this could represent some movement towards student empowerment, in as much as very clear and achievable indicators are given to the student about what they must do in order to complete the course adequately.

There is no theoretical reason why some specific behavioural objectives, and some more general criterion referenced objectives, should not be part of the negotiated contracts associated with the different ends and means definition of equity. However, these would generally be negotiated between student and teacher as part of the learning process, rather that imposed on students and teachers as predefined parts of the course.

In terms of its current usage in education, such negotiation would violate current practice and trends, which uses the criterion referenced outcomes, professionally developed and applied, as the true measure of achievement standard. Ironically, to the extent that the outcomes are inadequately defined, and thus confused, the gateway to incorporate such outcomes into the broad definition of equity becomes enlarged. That is, the outcomes may become differentially specific by negotiated discourse with particular students.

Because it denies hierarchy, however, this rarely happens. It is discouraging to see an assessment frame which seemed to hold promise for the empowerment of students now being used as an instrument of rigidity and conformity, as another meter to objectify disadvantage and enshrine privilege.

Hierarchy, equity and the Responsive frame

The Responsive frame contradicts hierarchy. Genuine negotiation implies symmetry of power relations. Openness in communication, the free flow of information in both directions, is not compatible with authority-subordinate power relations.

This would be true even if the power relations were reversed, and the student were to employ the tutor to teach. Dependency invariably inhibits truthfulness.

The Responsive frame is also contradictory both to the equal treatment and the equal ends definitions of equity. Responding to individuals in different ways is obviously not compatible with the equal treatment definition, and spontaneous generation of criteria, negotiated curricula and assessment descriptions, and obviously subjective responses, have little connection to common goals and end points.

This is not to say that some well-defined objectives might not be found acceptable and useful to particular students in describing what they wish to learn, and how they will know when they've learnt it. Nor that some other objectives may be so essential to a course that they are prescribed and proscribed in the beginning.

On the other hand the Responsive frame of assessment is quite compatible with the different ends and means definition of equity. This frame is, in fact, a necessary part of any educational processes that value diversity and freedom of students, and thus include this broad equity concept of fairness and justice.

Summary

The relationship of value to assessment mode becomes apparent. Certain definitions of equity, and certain assessment modes, are inherently contradictory to each other and to the power structures that contain them; as such, they will be seen, accurately and probably unconsciously, as potentially destabilising, and consequently be ignored, nullified, or corrupted into acceptability.

In the next chapter we look at the criteria of measuring instruments, and how these fit with the four frames for assessment.

Return to Table of Contents

Chapter 9: Instrumentation

Introduction

Assessments in the Responsive mode do not necessarily involve standards or measures. In this frame, assessors may be content to describe without measuring, to give feedback without judgment, to respond with blatant subjectivity.

However, in the political and technocratic world in which evaluation thrives, such "soft" assessments are scorned, and the claim to measure, to rank, and to compare to a standard is what gives status and power to the evaluation process. Sydenham (1979) points out that even in the physical sciences

In the social world, it is people, regardless of any particular label, who are subjugated.

Measurements in physics

To measure any quantity or quality in the physical world we use an instrument, and the instrument must be calibrated. To measure length we need a ruler, and on the ruler is the scale. To measure time we need a clock, and on the clock face is the scale in seconds. To measure current we need an ammeter, calibrated in amperes. The electricity meter measures electrical energy consumed and is calibrated in kilowatt hours.

To calibrate the instrument there are three requirements. The first relates to scale, the second to replicability, and the third to theory-practice bridging.

Whilst scales do not have to be linear (they may be logarithmic or indeed of any other mathematical or ordered function), the nature of the scale does need to be known if any sensible interpretation of the scale is to be made. I will discuss only linear scales here, as they are the simplest and the most common, keeping in mind that the general argument would apply to any other scale for which a mathematical function applies with which to interpret differences.

For a linear scale equal gaps represent equal quantities of the thing being measured. The gap between 3m and 4m is exactly the same as the gap between 6m and 7m. The period of time represented between 9.1 sec and 9.2 sec on the stop watch is identical to the period represented between readings of 12.8 sec and 12.9 sec. The 5 kw hr of electrical energy represented by the difference in meter readings or 39.4 and 44.4, is identical to the 5kw hr of electrical energy represented when the meter reading goes from 44.4 to 49.4. As we pay for the electrical energy that we use, we would want to be sure that this equation was true. We would want to be sure that equal differences on the scale equated to equal differences in energy consumption. And when measures are added we would want to be sure that the laws of arithmetic applied.

We would also want to be assured that our meter gave the same reading as any other meter. It wouldn't need to look the same, or even be constructed the same, but we would want to be certain that if other people used up the same amount of electrical energy that we did, their meters would also indicate that 5 kw hr had been used. So other meters and other occasions must give identical differences for the same energy consumption. Yesterday's 5 kw hr on one meter must be identical to tomorrow's 5 kW hr on another meter.

And finally, after being convinced that the scale was calibrated accurately and the results were replicable, we would want to be assured that the meter really was measuring electrical energy in the units described. We would not want to pay for 5 kW hr of electrical energy if we were only using three. If all the meters are over-reading we are all being equally ripped off, but we are still being ripped off.

To ensure this accuracy we would require comparison with some standard instrument, against which all others could be compared. Such a standard instrument would itself incorporate both the meaning and the value of the thing we are measuring. That is, the standard includes within its operation both the theory of its definition and the practice of its measurement. For example, a standard metre rule is both a practical measure of a metre, and incorporates the theory that equal distances along its length are of equal value. A standard Ammeter, designed to measure electrical current, incorporates within its operation both the numerical value of current marked on its scale, and, within its mechanism, the definition of the ampere as a particular force acting between two conductors a certain distance apart carrying electrical current. And our kilowatt-hour meter gives us a reading on the scale, and incorporates into its mechanism the definition of electrical power as the product of voltage, amperage, and time.

Strictly speaking, such instruments (as instruments), incorporate sub-standards rather than Standards; that is, because they are instruments, they necessarily incorporate an error, which in the cases cited is very small. Because the Standard, which is some fixed point on the scale, is by definition error free, it follows that the Standard must be defined in terms of some mathematical theory (or some replicable event that is more accurately measured than the instrument). That is, with theory or events which have been empirically shown to have specific linkages with other measurable aspects of the physical world.

The standard and the measure

At this point it seems important to clarify the fundamental difference between any standard, and the measurement of that standard, for it is in the failure to appreciate this fundamental distinction that much of the confusion (and manipulation and mis-information) about the measurement of human "ability" and "standards" is rooted.

The standard is arbitrary, and is completely accurate. It is not arbitrary in the sense that it is capricious or random. It is arbitrary in the sense that it is based on opinion, and is merely one of a very large number of standards that could have been chosen. However, once the standard is defined as the standard, then it is that exact value. The value of the standard measure is completely accurate not because it has been measured completely accurately; the value of the standard measure is completely accurate because it is a definition, and not a measurement (Sydenham, 1979, p26).

If now we wish to measure a particular thing, we may ask whether it is above or below the standard measure, and by how much. In order to do that we must measure it with an instrument of some kind, or make calculations that involve such measurements. And such measurements will always contain some error, for such is the nature of measurement, because measurements are made along a continuum, unlike counting, which occurs in discrete leaps. We may count the number of bricks, and may do this without error. But no two bricks will be of exactly the same weight. One will have a few more grains of sand or clay than another. And even if two were of exactly the same weight, we could never know that, for the instrument with which we weigh them also contains errors in its scale, in the calibration of that scale, and in the reading of the value of the scale. Two bricks for which we obtained equal weights could indeed be of different weights if measured on another scale of equal accuracy. And two bricks for which we obtained different weights could indeed be the same (within the order of accuracy of that measuring instrument) if measured on a scale of greater accuracy.

One of the party tricks used by educators and others who wish to defend their indefensible measurements is to give examples that reduce measurements to counting. Surely 18 out of 20 correct spelling is 80 percent! Surely number facts in addition or multiplication are either right or wrong! And then they stop. For in the whole field of education they can't think of any other examples where measurement may be so reduced to a counting procedure. Not to mention the sidestepping of the question, eighty percent of what?

The case of the digital watch

Increasingly, instruments use digital electronic mechanisms which use counting methods to give their scale readings. However, these jump from one number to the next, just as watches with visual dials jump forward in one second or tenth of second leaps. Time, however, does not jump forward in such leaps, but is measured on a continuum, as are most of the other quantities that we measure. So the upper limit of accuracy of such an instrument is the gap represented by the jump. The lower limit is much greater.

The interference effect

It is a truism of science, often conveniently forgotten, that any measuring instrument distorts the field it is intended to measure. This is obvious when we think about it. For the measuring instrument to operate, it has to interact - that is, interfere - with the field it is measuring. Newton's Third Law is a universal principal: every action has an equal and opposite reaction; if the field acts on the measuring instrument, then the measuring instrument simultaneously acts on the field.

The effect may be relatively small - a thermometer inserted into a large container of hot water will not much affect the temperature of the water, though it will affect it. However, a very cold thermometer inserted into a very small cup of warm water may cause the temperature to drop appreciatively. The temperature thus measured is not that of the hot water, but that of the water-thermometer system.

In this particular case, it is possible to estimate the imprecision caused by the measuring instrument, if we know the masses and specific heats of water and container and mercury and glass, and the temperature of the surrounding air and the time taken for the thermometer to give its highest reading and the rate of heat loss from the container. Then we may estimate the temperature of the water at the moment the thermometer was inserted. However, even in this simple case, it is necessary to use a theory that is itself, of necessity, subject to some imprecision.

Sometimes the instrument is permanently incorporated into the system, and can then be defined as part of the field. Our electricity meter is a case in point. It is a permanent part of the electrical fixtures in the home. Nevertheless, it does use up energy in its very operation, thus increasing the energy needed for the house. It does distort the field. And as we might expect, it is the consumer, and not the electricity company, who pays for the distortion.

So how big is the interference effect when a "test" is used to measure some human "attainment" or "ability"? How precise is the theory that links the measuring instrument to the thing it is supposedly measuring? And does the test introduce a small distortion into the field it is supposedly measuring, or is it of the same order of magnitude as the field? Are we putting a warm thermometer into the ocean, or into a little test tube of cold water?

Boundary conditions

Another fact of Science often conveniently forgotten is that the precision of the physical sciences - that is, their ability to obtain (almost) identical results in replicated experiments - is directly related to our ability to control the boundary conditions of the experiment: to prevent heat loss, to create a vacuum, to maintain a constant magnetic field, and so on. The precision of physics is specifically related to our ability to create a completely controlled (and hence artificial) environment in which to construct and conduct the experiment. The formulas of dynamics are very accurate in predicting the velocities of objects in free fall in a known gravity field in a vacuum. They are hopeless in predicting such velocities for a skydiver who jumps from a real aircraft in a real atmosphere. She will not reach the ground at the same time as a bunch of feathers or a lead ball thrown out at the same time, nor, luckily for her, at any time predicted by the formulas of simple dynamics. The point to note is that controlling the boundary conditions often produces an artificial environment which makes the data unusable in the "uncontrolled" world.

This excursion into elementary physics is occasioned not only by nostalgia, but by a desire to clarify some of the relationships between instrument precision and measurement precision in that most precise of sciences, and to point out that whilst precision in Physics certainly cannot be greater than that of the measuring instrument, and any calculation based on that measurement is limited by the empirical accuracy of the attendant theory, that in most cases these two variables are not the main limitation on replicable accuracy. It is rather the stability of boundary conditions, the physical scientist's ability to artificially freeze all other significant variables, that allows such precision, predicability and control in these sciences.

And this is the precise problem we face when we try to measure people. For the boundary condition for stable human behaviour (and all measurement of people, all assessments, all tests, all examinations, must elicit or refer to some form of behaviour), is a stable human mind. But the individual human organism is not a computer. It does not produce a unique response to the same situation, if for no other reason that the "same" situation never reoccurs. Perception and conception, and hence response, to "identical" situations invariably differ, as the variables that affect such reactions - attention, mood, focus, metabolic rate, tiredness, visualisations, imagination, memory, habit, divergence, growth etc. - come into play.

As Kyberg (1984) describes it:

So the very concept of a "true" measurement resides in the assumption of a stability and permanence in the characteristic being measured, and the boundary conditions of the measurement. Lack of these conditions does not represent so much an error of measurement, as a discrepancy with fundamental assumptions.

Where does the data come from?

Before dealing in more detail with the specific problems in measuring human ability, there is one more point to clarify. Where does the data come from? Where does it belong?

What Eisner is saying here is very important. The data, the measures, are not out there in the object being measured. They are measures that we have generated through a particular mechanism that includes the measuring instrument and the theory and some aspect or property of the thing being measured. Any claim to "scientific" truth involves a further implication that a similar mechanism would produce similar data on another occasion with the same person. Or more accurately, with the person that person has now become.

So the temperature is not only some aspect of the object being measured; it is also and equally a meaning generated by a certain way of construing the world (the theory), and a certain way of interacting with it (the mechanism which includes certain actions with instrument and object). As Pawson (1989) expresses it, the only alternative is "to retain the notion of an observable realm that is independent of us yet knowable, . . . (and) to propose some automatic, pre-established harmony between subject, language and world"(p61).

In like manner, if we are able to measure some aspect of a person called their ability, we are not measuring something they have. We are generating data that is also determined by the mechanism of the instrument - person interaction, as well as by a certain way we, the assessors, have of construing the world. In other words, we ask them to live in our little experimental world for a time, and make a measure in that world. To pin the label on them apart from that world is to misrepresent the experiment: The data, the label, belongs not to them, but to the whole theory-experiment-instrument-object interaction.

Measuring human ability

The rather detailed account of the properties that measuring instruments must have if they are to be usefully used in the study of the physical world enables us to look more adequately at the measurements being used in the study of human ability or human attainment. We might expect such instruments also to incorporate the three same necessary elements: a generally acceptable theory that enables the gap between theory and practical measurement to be bridged, in which a standard measure is defined; an instrument that is itself replicable in terms of the theory, and gives replicable results when measuring the same thing on different occasions; and a scale on which equal differences either represent equal "ability" differences, or can be translated into some meaningful comparison by a known mathematical relationship. This last becomes particularly important if we wish to use it to make a categorisation, or be added to some other measure.

Standards and standards

Before examining how the Judge, General, Specific and Responsive frames for assessment stand up in relation to these three elements, I want to clarify the meaning of the word "standard" in relation to human products. This "standard" relates to a point on a scale, to a point below which the product is unacceptable. The standard thus indicates the lowest limit of acceptability. It requires a scale to define it.

This "standard" is utterly different to the "Standard" which is the basis of the scale, and hence of the measures made by the scale. This "Standard" defines a difference between points on the scale, and can be used therefore to check the replicability of instruments. So we have a "Standard" metre length, a "Standard" second of time, a "Standard" kilogram of mass. I have (arbitrarily) differentiated this Standard with a capital S. Such Standards are useless unless measuring instruments of great accuracy are available to sub-divide and expand the scale embedded in the Standard. However, the specification of any Standard does not guarantee the existence of a suitable measuring instrument (Sydenham, 1979, p26).

The tendency we have to attribute guilt by association is well known. We are less wary of the tendency to attribute innocence by association. Our Standards of length and time are immensely accurate, as any Standard that defines a scale must be. Indeed, Standards of this sort are infinitely accurate because they are definitions and not measurements. The sub-Standards do involve measurement. And as the sub-Standards also provide bases for scales, the measurements they make must be very accurate and precise. We tend to associate similar accuracy of measurement to those quite different "standards" that are used to describe minimum acceptability.

Most industry product "standards" of minimum acceptability are based on criteria for which very accurate measurements can be made. That is, we can measure very accurately whether our product is minutely above or below the stated standard. And that tends to make us forget that the standard itself is not a measurement but is a definition, and is arbitrary. Any amount of a particular additive to food could be harmful to a particular person. All exposure to radiation, even background radiation, has an effect on living organisms. Any bridge will collapse under some particular conditions. Product standards are always statements about a compromise. They represent the arbitrary point at which safety, conservation, style, cost, expediency and whatever strike an uneasy, indeterminate, and hence arbitrary balance. At which point they assume a solidity and stability that denies and contradicts their genesis.

Any standard of acceptability is a political entity, as much in its production as in its enforcement. The myth of certainty that surrounds measures of people is achieved partly by its association with the Standard that defines accurate scales, and with the standard that is a definition of acceptability. As well as the standard we salute as the symbol of authority, as referred to in chapter 6.

Judge's frame

Whilst the Judge often uses a student's written work, in assignment or tests, as a basis for measurement, the Judge would not see the test as an instrument. Nor would he claim to make a measurement. What is written is merely a vehicle for showing him what the student is capable of. The Judge would claim to be able to use any such example as a basis for indicating the level that the student had attained. The Judge is not even particularly concerned to have a sample, random or otherwise. Any example, according to the Judge, can be judged according to its relation to the standard.

In scientific terms, the cognition of the Judge is the instrument, and incorporates the Standard, the scale, the theory-practice gap, the standard of acceptability, as well as the actual measurement, all within its own internal mechanism. Putting it more bluntly, the Judge simply does not operate on a scientific paradigm. Rather the Judge is a mystic who claims to "know" the definition of standard, rather as one may "know" the presence of God. A student's level of attainment may then also be "known" and hence judged accurately, through the union of his/her own consciousness and that of the person being assessed, the example of the work judged being the medium through which this communion occurs, rather in the manner in which tea-leaves activate the astral consciousness of the psychic. Such a process is sometimes conceptualised and rationalised by considering the permutations of such value imponderables as style and form, understanding and creativity, texture and design, understanding of the field, or whatever. Many, if not most judges, would admit however that such variables were used to justify their intuitive judgments, rather than to logically develop their proofs.

From the point of view of the scientific paradigm, the work of the Judge is aesthetic rather than scientific. As such, it belongs logically to the Responsive frame with all the limits and advantages of the overt subjectivity of that frame. Creative reflections on their work by others can be of great value to a student's learning. However, when given in the form of absolute judgments rather than helpful feedback, such reflections are more likely to stifle learning than to expand it, more likely to inhibit creativity than encourage it, more directed to conformity than diversity.

What stops such classification into the Responsive frame is the refusal of the Judge to admit such idiosyncratic subjectivity, and to insist on the truth and objectivity of his judgments as measures of human performance or ability, by invoking the ideology of the absolute standard and the expert judge, and assuming, in both senses of that word, a state of mystical communion.

More recent post-modern conceptions of the Judge's frame use the notion of the interpretative community to defend the position of the Judge. Here quality is determined by a discourse embedded in the language of the field, and various criteria or aspects of quality may be so discussed. However, despite the acceptance within the community of the ephemeralness of the notions it produces, the end result is still the categorisation of the product and/or the student; a solid dichotomous categorisation that denies the tentativeness of its genesis, and, certainly outside that community, and I suspect also within it, is not regarded as a problematic (Fish, 1980).

General frame

The General frame pays considerable attention to problems of scale and replicability, and the theory-practice gap. Theoretically (though almost never in practice) it uses random sampling theory and practice, and assumptions about the distribution of attainment, to produce an instrument (a test), define a scale (normalised score), and estimate replicability (standard error or correlation). In terms of "ability" measures various standards can also be defined in this model to comprise certain grade levels, in terms of percentiles of defined populations.

Now this is more or less what "standardised" tests do. In my view they vastly underestimate the error, both in its theoretical definition, as well as in its representation (or more accurately its non-representation) to student and faculty. Some specific details of this are given in Chapter 15 on the psychometric fudge. Rarely do the instruments satisfy the requirements of theory (random selection of items), nor do the populations on the basis of which they are calibrated (random selection of the population). Even so, they do tend to satisfy some of general requirements for a measuring instrument, as required by the physical sciences, even though the errors in these instruments, if made explicit in public knowledge, would make them useless for the purpose for which they are designed.

There are however three more fundamental sticky points, points at which the whole exercise becomes very suspect, or unrealistic. The first is inbuilt, and concerns the assumption about normal distribution of performance (or indeed any other assumption that might replace it) built into the theory. There is absolutely no reason to believe that in any area of educational activity the end point should be represented by a normal distribution (which is the same shape as a random distribution) of attainment. In fact, the better the educational environment, the more likely we are to obtain a very skewed, lop-sided, distribution of attainment.

The second occurs when the scores, which are defined in terms of the distribution, are presumed to relate to some "standard of competence" for an individual student. This latter represents an error in logical typing, but might be more truthfully described as a semantic confidence trick.

Perhaps the most blatant example of this is the distribution grades that are labelled A B C D F. These grades may be defined in terms of percentile distributions, so that A represents the top 5 percent of the rank order of students (or whatever other arbitrary percentage is chosen), B the next 20 percent, and so on. Logically then, F represents the last 5 or 10 percent or whatever. So why not E? Because F also stands for "fail," a statement about competence and not distribution. And historically, as we know, A and B have connotations of excellence that C does not have, though there is nothing in the distribution that implies either that A is an excellent performance, nor that C is a mediocre performance. For example, if a group of professional sprinters throw the javelin and are then graded in terms of their rank order, we would not expect those obtaining an A to have reached the Olympic "standard". On the other hand the person who runs last in the Olympic 800 metres final is hardly a mediocre runner, or a failure.

For even if we except the notion of a "normal" distribution, the sticky question still remains: a normal distribution of which group? All the people in the world? All the educated people? All the people still at school? All the fifty year olds? All the people at a particular grade level? In a school? In a city? In a country? Without this detailed information the "standards" cannot be given a meaning. And even with them, they can be given no meaning other than that defined for them. That is, their meanings can only relate to distribution, and not to competence.

Even with such information about the nature of the sample population, there is, and can be, no formula, no equation of equivalence, between grades defined by distribution on a rank order, and some pre-specified level of attainment of an individual student (Airasian, 1979, p42; Jaeger, 1980, p64; Glass, 1978; Levin, 1978, p314; Burton, 1978, p263; etc. ect. ect.).

In addition, the differences in logical type in attempting to make linear measures of complex qualities generate paradox and confusion and hence strong emotion and unresolvable debate (See Chapter 12). This makes the topic utterly suitable for creative endeavour and satirical humour, but impossible for scientific measurement.

The third point is more fundamental, and may well make the other two points trivial. There is no Standard against which the scale can be calibrated. There is no theory that enables a definition of some point on the scale to be distinguished, against which the scale might be calibrated, along with other scales purporting to measure the same thing. The test scale floats freely in space, relating solely to its own assumptions with no Standard rope to bind it to the earth. What we have here is not a scientific instrument, but a very suspect ordinal scale pretending to derive from a scientific measurement.

Specific frame

In the "pure" Specific frame, a person's "ability" or "performance" or "attainment" is reduced to a finite number of specific behaviours, for each of which a "standard" is clearly defined. Thus we are, in theory, able to specify exactly which "objectives" have been achieved to the specified "standard". The notion of scale, Standard, and measuring instrument is (apparently) sidestepped by postulating a dichotomous variable, requiring not a scale, but rather an on-off switch, to categorise its measure. We shall come back to this in Chapter 11, where it is argued that all categorisations infer measurements.

However, in most areas of human endeavour such reductionism to specific behaviours results in trivialisation of the task. Further, specification of the "standard," even in such a narrow and specific thing, is still very difficult in most cases, as the measurement instrument does not exist, and it is finally fallible human judgment which in practice must decide whether the standard has been achieved for each objective. Further, the basic assumption is erroneous; the variable being measured is continuous, not dichotomous, so the measurement error still exists, disguised though it might be. We are back again to the Responsive frame, requiring a subjective decision, which is covered up by pretending to be the Judge's frame, requiring an unambiguous omnipotent objective decision, which is in turn covered up by pretending to be an example of unambiguous standard in the Specific frame, derived from a definition of standard which pretends to be dichotomous and pretends to be nonarbitrary.

To further confuse the issue, what often now happens is that specific information about which particular objectives have been achieved is lost when measurement is reduced to counting, and the number of objectives achieved is the only information recorded. This creates the illusion of exactness and error-free information by disguising the fact that the exactness of the "standards" of individual objectives is, in practice, illusory.

Responsive frame

In the Responsive frame the person's work, or inferences about the person's capacity or ability, are described but not measured. Further, these responses are ideally owned by the responder, and not projected onto the producer, or the producer's work.

They may describe how the person's performance relates to certain criteria, how then the performance might be improved, and to what extent, in terms of such criteria, and in the opinion of the responded, success has been achieved.

The responded may also offer some opinion about whether the work of the person being assessed is of inferior or superior quality, or whether they are skilled enough to practice in a certain field of work. However, again, this does not purport to be a measurement of some clearly defined standard, but merely the informed view of a particular person who for some reason or another has views worthy of hearing. As Stake describes it:

But note how quickly Stake modifies the insight of his first sentence with the caution of the second. In whose interest is this emphasis on quietness? Why this concern to legitimate resistance rather then stridently call for reform? Who might hear strident voices, that quieter ones may not discern? And whose voices go unheard in the quest for quality, and the demand for categorisation?

And note also the very narrow gap between offering an opinion on whether the performance is adequate for some purpose, and categorising the student. We are here at the very edge of the Judge's frame of reference, a boundary crossed over as soon as the categorisation is made.

Summary

In this chapter we have looked at the invariances required in events involving measuring instruments if such events are to have credibility. In particular the notion of a Standard that theoretically defines the scale, and how that is not to be confused with a standard of acceptability, which is to be measured by the instrument, and which requires a scale in order to be located. We also noted the importance of the specification of boundary conditions and interference effects, and that the price of invariance and tight theory - practice links was artificiality.

The various assessment modes were then analysed in terms of their instrumental error. All were found to be invalid, on the grounds of not satisfying the conditions of adequate instrumentation.

Return to Table of Contents

Chapter 10: Comparability

Synopsis

In this Chapter I examine the notion of comparability as it applies to the assessment process. Any rank ordering of students, any adding of marks on examinations, any addition across subjects, assumes that comparisons can indeed be made.

The fundamental distinction between more and less, and better and worse, is first elucidated, and this is linked with ideas of uni- and multi- dimensionality and notions of doing or having. This analysis is then applied to ideas of traits, abilities, and skills, and their supposed measurement in tests and examinations. Some fundamental confusions are exposed.

The discussion then moves to what meaning if any can be given to the result when marks or grades are added, how loadings on final rank orders are affected by spread of marks, and how differential privileging of sub-groups occurs with different intercorrelations. Finally, it is contended that for individual students the privileging is non-predicable, and the total score thus meaningless.

Goal kicking skills

> *George!*
>
> *Yes coach?*
>
> *You know why we've lost the last six games?*
>
> *The other teams were better?*
>
> *Bad kicking, George. Bad kicking. And with six in a row, someone's got to go.*
>
> *Gee coach, that's really poetic.*
>
> *Yeah George, and you're really pathetic. Anyway, do some tests and get me a team ranking on best to worst on goal kicking skill.*
>
> *No worries, coach. Goal kicking skills, you said?*
>
> *That's what I said. Get me a best to worst ranking on goal kicking.*
>
> *What particular aspects of goal kicking, coach?*
>
> *You're the trainer, George. How far they can kick. How*

straight they can kick.

Anything else?

Jeez, what do I pay you for? Set kicks, kicks on the run, and snaps. That ought to do for a start.

No worries, coach. I'll work out some tests for each of those and give you a list in a coupla days.

(Two days later).

Here you are, Coach. Here's the list. I've ranked twenty five of them in order of merit on goal kicking skills.

That's great, George. Just what I wanted. Let's have a look at this. Harvey's on top of the list. How many goals has he kicked this season?

None, coach. He's been playing in the back pocket.

Look where you've got Shonker. Twentieth. He's the bloody full forward. He's booted a hundred goals this season already.

Yeah! well, he's missed two hundred.

So he's missed two hundred. He's still booted four times as many as anyone else.

That's because he has ten times as many possessions as anyone else. You didn't ask me about that. You just asked me about goal kicking skills.

Yeah, OK. So who's the longest kick?

Can't tell you that. It got lost in the data.

Who's the most accurate on set shots over 50 metres?

Got lost in the data.

Who's the best snap shooter. No, don't tell me. Got lost in the data.

Hate to tell you, coach, but I think this list is a load of shit.

You can say that again. Who was the idiot who did it?

The idiot who did what some other idiot told him to do.

Better or more?

Fundamental to the process of arranging orders of merit is the notion of comparability. As we have seen, the notion of standard implies the notion of order of merit, which implies the notion of more or less, better or worse. For such notions to have a meaning, they must refer to some aspect, some property that is being compared, that is presumably being measured.

Regardless, the first paragraph slid past a fundamental distinction: "more or less" is not the same as "better or worse": More or less are terms related to counting, to mathematics, to scales and measurements. They are loaded with notions of objectivity, and solicit entry to the quantitative world; better or worse are terms related to value, to goodness. They are permeated with the aura of subjectivity, and are related to the qualitative world, the world of valuing. The concepts are in different domains of discourse. If the criteria is size, then two people may be compared as being more or less heavy; or their weights may be compared in terms of better or worse in regard to health. But the two ratings are unrelated. Or if the criteria is emotionality, we may rate people in terms of whether they are more or less emotional; or we may rate them in terms of the appropriateness or productiveness or empathic clarity of their emotionality. Again the two ratings are conceptually unrelated. Or so it would seem.

What is the essence of this difference? For when we tried to explain what we meant by better, we used words like healthy, productive, empathic, clarity: and the interesting thing is that we may use more or less with any of these words, even though we started off in the better or worse category. And we may also ask of each of these new criteria whether they are better or worse; in this case questions preempted in the predominant paradigm because value judgments of better are already built into the words chosen to describe the criteria.

So what is the essence of the difference? In relation to aspects like size or emotion or clarity, when we ask the question more or less we are asking about intensity, about how much or how many. We are referring to the aspect in isolation from its environment. The event that produces the judgment about more or less involves our sensory relation to that aspect independent of other aspects. More or less questions are answered by focussing on the aspect and on no others. More or less questions are directly answerable. The answer may be incorrect, but such a statement in itself implies that there is a correct answer. More or less has only one meaning in relation to a particular aspect. They can't be more and less at the same time, so the question is convergent, and presupposes a world in which there is a true answer to the question. So logically more or less implies a uni-dimensional aspect, a world of transitive and asymmetric relations (Lorge, 1951, p548).

On the other hand, when we ask the question better or worse, we have to ask another question, In what way better or worse? Because something may be better in some ways and worse in others. Better or worse in what aspects? Or better according to whom? Or better under what conditions? And when we nominate those aspects we can ask of them two questions about any comparison; more or less, or better or worse. And so on. Essentially better or worse implies multi-dimensionality in the aspect under consideration.

What does all this mean? Very simply, when we ask the question more or less there are no further questions to ask. We move straight on to the answer. In other words, more or less questions define the end of discourse; they are a direct invitation to a judgment; they are the signal to stop thinking, and act; and incidentally and significantly, to accept the judgment, which comes after the thinking has stopped.

But the question better or worse logically invites more questions about the first criteria. In what way better or worse? Which introduces more aspects, particular aspects selected in most cases from a much larger set of possibilities. For there are as many aspects as our conceptual imagination may produce (Lorge, 1951, p536). Yet the original aspect is reduced, even as more precision is generated by defining aspects; and as more aspects are conceived, the potential disparities of the judgments concerning them increase. And then for each of those aspects: More or less? Better or worse? And again, the additional questions about positioning and context are generated. So better or worse questions encourage further discourse, and further thought.

All this is not to deny that the power relations in which such discourse is embedded may dictate that the answer to the question better or worse be given at any time and be accepted without further thought. But that in no way invalidates the additional logical questions that the aspect implicitly generates.

Having and doing and being

It is obvious, but important, to make the point that whole entities (holons) cannot be directly compared in terms of more or less, only aspects of them (Jones, 1971, p335). One dog cannot be more than another dog. Nor can a stone be more than another stone, nor a stone be more than a dog.

In like manner dogs and stones cannot logically be compared in terms of better or worse, for such a claim is meaningless without a response to the question "in what way better?" A dog cannot be better than another dog. In terms of dogginess, dogs are equally doggy; they are equal by definition, as being classified as dogs. Likewise with stones. And dogs and stones cannot be compared as entities because they are in different classes. It follows that the very act of classifying whole entities (into classes) logically invalidates any comparisons within or between the entities that comprise them. Classes of course can be compared in terms of the numbers of elements they contain, but this is a different matter.

Two people are being compared in terms of the relative merit of some task. In terms of doing, we may say that one person does it better than the other. This is a statement about relative merit. Or we may say that one person does it more than the other. This is a statement about relative frequency, and not of relative merit. You may drive a car badly many times.

In terms of having, we may say that one person has more of something than the other. This may claim to account for the greater merit. It is essentially a statement about the comparative number of elements in a class. But we would not account for a difference in merit by saying that one person had that something better than the other. Such a statement refers to the whole class and whole classes cannot be compared except by numbers of elements.

So in terms of relative merit, the question of more implies a different mode of description, a different ontology, than does the question of better: Better or worse is a comparison of what people do under certain conditions, made by some person; more or less is a comparison of what people have, or are alleged to have. As such it is logically independent of any contextual or positioning variables. One begins to see the simplistic delusion generated by mathematical modelling.

Logically then better or worse questions cannot be answered definitively until they are reduced to a criteria which comprises a class in which the question better or worse is reduced to the question more or less. Logical here means relations that are transitive and asymmetric.

Pragmatically, better or worse questions can be answered whenever the criteria are sufficiently understood (implicitly or explicitly) to allow consensual subjectivities of judges to give similar answers. However, as we have indicated earlier, such criteria are multi-dimensional. And as is evident from the conversation that began this chapter, little if any meaning can be given to a uni-dimensional description of this multi-dimensional entity in terms of their uni-dimensional elements. As we shall see later, one meaning of such a comparison is dependent on the relative loadings of the different dimensions.

Politically, of course, better or worse questions are answered whenever someone with sufficient status or power gives a decision.

Comparing people

It follows that to compare people, whole people, we may compare either some parts that comprise them, or some wholes of which they are parts. If we look at the parts that comprise them, we may look at the person's elements or internal processes; if we look at the wholes of which they are parts, we may examine the person's functions and relations in the wider environment or community, or at the cultural meanings in which their thoughts and actions are embedded (Wilbur, 1996).

Let us compare two people in terms of their relative merit in Physics. We are particularly interested in their relative achievement in a particular course of study at year 12 level. Such a course has a range of content and objectives and involves practical and cognitive operations of varying complexities.

We are obviously in a multi-dimensional world, in which at this stage more or less questions are meaningless. Further, any logical answer to the better or worse question is going to depend on the details of the answer to the prior question: In what way better? What particular aspects? Under what particular conditions? In whose opinion?

And if we intend to give a meaning as well as an answer to a multi-dimensional comparison, what are the relative loadings of each aspect in the final judgment?

Of course, we could simply ask the teacher who taught them, who is better? And the teacher might give a judgment. But in making sense of that judgment in terms of the original question, the implicit questions still hang there; in what way better? So after the judgment, the teacher must logically justify the decision on the basis of criteria; and if one is not better on all possible criteria, then the question of how the criteria are loaded to obtain the final criteria is relevant.

So, either prior to or after the judgment, how might the discourse progress?

> *In what way is she better?*
>
> *She knows more facts.*
>
> *Is that all?*
>
> *No. she's better at solving problems?*
>
> *In what way better?*
>
> *She gets more complex problems right?*
>
> *Does she get more simple problems right?*
>
> *No, he gets more simple problems right?*

In what ways is he better?

He is more careful, he makes less mistakes.

And so on , and so on. And if we are dealing with twenty or thirty persons, it is clear that different criteria of comparison are possible for each pair, and there is no reason to believe therefore that there would emerge any final rank order of merit, for on the basis of different criteria of comparison, A could be better than B on criteria 1, B could be better than C on criteria 2, and C could be better than A on criteria 3. This is an empirically inevitable consequence of multi-dimensionality. It is inevitable because only when every criterion correlates unity with every other criteria will ranking invariance occur. And in that situation we are, by definition, in a uni-dimensional situation. It is the reason that psychometricians fantasise unmeasurable but uni-dimensional true scores.

Viewed from this perspective, it becomes clear that the more specific, limited and applicable to all comparisons the criteria become, the more possible it is to finally reduce such aspects to those answerable by more or less, the more possible it is to produce an invariant ranking, and meaning (in terms of explicit loadings) for the meaning of the original comparison. However, such meaning is at the expense of initially reducing and finally confusing the meaning of the original comparison. Another example of the essential contradiction between reliability and validity.

Traits, abilities and skills

A trait or an ability is a thing that a person has. A trait is a hypothetical entity, an abstract attachment, a comparative label, that is used to explain differences in what people do in terms of something that they have. A trait is described not so much as a performance as a potential performance, as a sort of template of the performance that might emerge under ideal conditions, whatever that may mean; a morphic field that predates performance. This magical property of a trait makes it forever immune to particular environmental conditions, which may indeed influence particular performances, but leave the trait, securely protected within the person, unsullied and unmoved, firmly fixing individual merit in correct relative position in the grand order of things.

A skill is a much more difficult ball of wool to untangle. A skill is something you have, like a verbal reasoning skill. On the other hand, a skill is normally exhibited as something you do, like playing a musical instrument or tennis. And you can have more skill but maybe not better skill (skill here is used as a holon). On the other hand, you can have more skills or better skills, and these two meanings are different, as with the goal kicking skills referred to earlier. Better skills here appears to have more to do with a particular selection of skills relevant to a particular context. Then again, skill seems to refer at times to a particular standard in a more-less or better-worse ranking; unskilled refers to rankings below the standard. It is clear from all this that the word skill is a very useful word to have in any discourse that wishes to imply precision even whilst it multiplies confusion. Norris (1991) notes a similar confusion in the notion of outcomes:

> The precise specification of performance or outcomes rests on and leads to a mistaken view of both education and knowledge. Mistaken because there is a fundamental contradiction between the autonomy needed to act in the face of change and situational uncertainty and the predictability inherent in the specification of outcomes (p335).

The world of objective tests

Objective tests, which often claim to be value free, necessarily do not ask better or worse questions. The whole operation is contrived so that only more or less questions are asked and answered. Further, they necessarily deal with what people have, not with what they do. Thus it is not so much a desire to deceive that drives the psychometrician to imagine constructs such as ability or traits or skills, but a logical necessity of the world they have constructed.

For it follows that if there is to be an answer, rather than a multitude of answers, to a comparison of two people, it is essential that the question better or worse never be asked, and all comparisons be reduced to the question more or less.

So the world of objectives tests, like the world of chess, and the world of mathematics generally, is certainly internally logical. Whether it relates to anything that actual people do in the world, apart from answering objective tests, or playing chess or mathematics, is another question.

The world of public examinations

Examinations live in far more dangerous territory. The constructors and markers of examinations are far less isolated from the front line of educational activity than are test writers. Their language is less precise, their pragmatism more up -front, their compromises and contradictions more overt. So they are far more likely to slide uneasily between concepts of better or worse, and of more or less, according to the pragmatics of phases of the assessment.

Consider the marking of essays. Whilst guidelines for marking may be given, ultimately notions of better or worse must be utilised by examiners in deciding what mark to give. Such guidelines are designed to circumscribe the answers to the question "what aspects?," to limit variability in the question "who says it's better?," and hopefully bypass entirely the question of the effects of the conditions on the essay's production.

So in stage one, the answer to the question of "better or worse," which establishes the ranking of students on a particular question, is used to determine the answer to the question "more or less," which is the mark given. Now the marks are added to give a total score, which is then interpreted as being better or worse according to whether it is more or less. Finally, if the grades are not distributed statistically, someone must look at whole papers around the grade boundaries to decide which are in their opinion better than the standard that defines the boundary, and which are worse.

Now, it is clear that this procedure only makes sense if the notion of better or worse, and the notion or more or less, are synonymous, within the series of events that comprise the examination. In other words, if better means more within the context of the examination. Practically, this makes it now impossible to untangle the interaction between the two notions, or deal with the complexities involved when multi-dimensional aspects are mapped onto uni-dimensional scales.

It is not my intention to suggest a solution. It is my intention to establish a confusion, and to note that such confusions must invariably lead to more invalidity and uncertainly about what is being described here. In other words, here we have another, crucial and fundamental, source of error.

We are tapping here one of the distinctions between quantity and quality, two concepts often fused together in discourse on measurement and evaluation. At this point it is sufficient to note that big is not necessarily better; getting more sums correct than somebody else does not necessarily make you better at mathematics: nor does getting more spellings correct make you better at writing, or getting more multiple choice questions correct on a philosophy test make you better at philosophy, or a better philosopher. To suggest otherwise is perpetrate a category confusion. The matters raised in this paragraph are further elucidated in Chapter 12.

What can be compared? What can be added?

So in terms of "more or less" we can compare any events that have a common aspect, that have a criteria on the basis of which we can rank them in terms of having more or less of that common aspect. A criteria, that is, that can be considered uni-dimensional.

Two questions then arise, which are fundamental to the whole notion of testing, examining and credentialling. The first question is, what happens when we add measures or ranks that relate to the same aspect? The second question is, what happens when we add measures or ranks that relate to different aspects?

Let's compare swimming pools in terms of two aspects that are comparable in terms of the same measurement units, a claim incidentally we could rarely make in the human measurement field; we could compare the pools in terms of length, or in terms of depth. In both cases they may be measured accurately (to within one millimetre) in metres. Now we could obviously compare our pools in terms of length, and we could compare them in terms of depth. The question is, could we use these criteria to obtain a single measure in terms of which they could be compared? This is in many ways an ideal situation; we have an accurate scale and measuring device, and our two aspects can be accurately compared on the same scale. So we could add the measure of length and the measure of depth. But what would it mean?

We could classify swimming pools uni-dimensionally in terms of the sum of their length and their depth. In terms of the initial components we have now lost any meaning, but the process (the addition) does enable us to imply another meaning; in this total positioning length and depth were equally valued, because we added the two measurements together, each with a loading of one. Or so it would simply appear. But things are not always what they seem and in this instance this would be an erroneous inference.

The relative valuing of the two components may be looked at in two ways; in terms of absolute value of the combined measure, or in terms of the influence on the rank order of the combined measure. Let's look at the absolute measures first.

If the depths of the pools varied from 1 metre to 2 metre, whilst the lengths varied from 10 metre to 100 metre, magnitude of the addition would be almost entirely defined by the length measurement. Alternatively, if the lengths of the pools were all between 15 metre and 16 metre, and the depths varied from 1 metre to 5 metre, then again the length would contribute most to the total measure.

However, in the second case the final rank order of the total measures would be most influenced by the depth measurement, which has a bigger range. So whilst the loadings for absolute values of the sum of measures are determined by the absolute values of the components, (which could statistically be characterised by their mean value, if we wanted to lose a lot of information), the loadings for determining the final rank orders are determined by the standard deviations of each component (Guilford, 1965, p424).

In this situation, the rank ordering of the total can be given a (process rather than content) meaning in terms of the relative valuing of the two components; and that valuing is implicitly determined by the standard deviations of their measures. We may adjust this by loading one of the measures. For example, a diver may greatly value depth over length in his pool, so may want the addition to mirror that valuing. So the diver may want to load the depth scores (by multiplying by a certain number) so that the standard deviation of the (loaded) depth measure (before addition), is 5 times that of the length measure. On the other hand, a long distance swimmer may want the two dimensions loaded the other way. In both cases the specific loadings are arbitrary, and in both cases they are related to function. And in both cases the final measure has no meaning other than that attributable to the relative contribution of each component to the final measure. (Of course, in this case the addition was completely unnecessary to the function; it would have been more rational for the diver to specify a minimum depth and minimum length, and for the long distance swimmer to do likewise; but that would have left us with no single variable with which to compare pools. And as mentioned elsewhere in this thesis, that may be the whole point of the exercise).

Let me generalise a little from this very simple case;

- 1. Any measure implies a ranking. Rankings imply transitive and asymmetric relations.
- 2. Rankings of a single aspect have a meaning, in terms of relative size or intensity of that aspect, which we can specify as more or less, and hence by numbers.
- 3. Rankings of different aspects may be added, but the addition has no meaning in terms of either of the aspects taken separately; the addition can be given a meaning in terms of the relative contribution of the two aspects to the total.
- 4. The relative contribution to ranking is determined by the loadings, equal to standard deviation multiplied by an arbitrary number.

The effect of correlations on loading

Let's go back to test and examination scores. We have three sets of scores (L, M, N) for the same group of people. The scores have the same standard deviation. We wish to add them to get a total score. Our theory tells us that they will have equal loadings on the final score.

Assume L and M scores correlate zero. Then when we add the L scores to the M scores, rank orders of both are changed, and it looks as though they contribute equally in determining the final rank order.

Assume M and N scores correlate one. Now when we add the N scores to the M scores the rank order of the M scores is unchanged. We could argue that the N scores have contributed nothing to the rank final order.

But then, if we add the M scores to the N scores, we could argue that the M scores contributed nothing to the rank order. A paradox. It is not necessary to resolve the paradox to realise that in this case the loading is determined by what is being added to.

It is also very clear that the final rank orders are very different in the two cases of zero correlation and unity correlation.

Regardless of the loadings (statistically determined by the standard deviations), different students have been privileged in the two situations described. In the uncorrelated (r = 0) groups, no particular group of the M score group is being privileged, or under-privileged, by the addition. However, in the perfectly correlated groups (r=1), the students who do better in M scores are all privileged when the scores are added, and the students who do worse do worser when the scores are added. This is in addition to the fact that the standard deviation of the composite score is 1.4 times greater in the case of the perfectly correlated group, giving it just that much extra loading as a composite when compared to the other total (Guilford, 1965, p418).

So what does all this mean when both L and N scores are added to the M scores to obtain a single rank order? The L and N scores both have equal loadings to the M scores; but this is a group phenomenon, and tells us little about individual students or sub-groups of students. We have seen that the L score loadings are more or less equally distributed across the M scores, but the N scores have privileged the top sub-group (according to M scores) and down-graded (with respect to the total score) the bottom sub-group. By interpolation we can see that this phenomenon will have a differential effect over the whole range of possible correlations and will be greater as the correlation with the scores added to increases.

In addition, to the extent that the means of the L and N scores are different, to that extent will the addition scores generally privilege the group with the higher mean.

It is clear that the statistical notion that relative standard deviations determine loadings is a vast oversimplification when applied to complex comparison situations.

Comparability, true score, and error

Here we have presented, in very simple form, one of the dilemmas of public examiners who must cope with adding different scores, from different subjects, or from the same subject marked internally and externally, and end up with some final rank order of marks because someone has said this is what they must do.

I have argued that such a total score can have no meaning other than that inherent in the loadings attributable to each component added; and I have shown that whilst the loadings of the whole group from any one school may be controlled through controlling the standard deviation of the marks, the correlations of the score with the score added to will influenced the subgroups which are over or under privileged by the addition.

There is another paradox evident in the conclusion, especially in regard to internal-external scores. To expose the paradox two further facts need to known.

Firstly, the rationale for internal assessment is that something different (broader, deeper, more complex, more varied) is measured by the internal assessment. Secondly, we can assume that in most public examinations some twenty to forty percent of students will be deemed to have failed, and to that extent the rank orders of their final scores are irrelevant in respect to the grades of those who pass; so the pragmatic teacher might argue that to underprivilege students who will fail anyway "does not matter."

In such a situation, it is rational (if somewhat inhuman) for schools to aim for maximum correlations with the external examination in order to privilege those who will most benefit from such privilege (that is, the best students). However, in order to do this they must invalidate the internal examination; for such an examination is surely more valid the less it correlates with the external scores, because it is supposed to be measuring something different. In short, the price of success is invalidity.

The middle way

> *That's all very well for the front runners, but most of the kids I teach are more middle of the road. I just want to get as many as possible past the cut-off point for entry to University or TAFE.*
>
> *Well, you've got a different problem then. You want to maximise opportunity for the middle group, not the top group.*
>
> *I suppose you could put it that way. So how do I do that?*
>
> *Easy. Just take out that middle slab of students and put them at the top of the rankings.*
>
> *Just like that?*

Just like that!

But isn't that unethical? Doesn't that make the whole examination invalid?

Sure. But as I've explained, it's invalid already because of what many schools are doing for their top students.

Are they really aware of what they are doing?

What's the difference. I don't accept the view that in this case bliss in ignorance makes the position less unethical. It certainly doesn't make the practice less invalidating, or the errors less significant.

When equal loadings are unequal

I have shown how equal loadings for a group may take on different shapes according to the correlations. Equal loadings for a group does not in practice mean equal loadings for all subgroups of that group. And in terms of individual students it doesn't have any particular meaning.

The question then arises, does equal loading for the whole group of students mean equal loadings for each separate school? Surely some school groups are really better than other school groups so should be differentially loaded? Some school groups might have higher means, and some may have larger or smaller standard deviations in the sets of marks that indicate their comparative attainments. And these might mirror differences in intrinsic ability, whatever that means, or might be a function of very good, or very bad, teaching, whatever that means. But if such students are tested internally, how would we know about their differential potential, or their differential attainment, as distinct from differential testing effects? And especially how would we know if they study and emphasise different things, and value different criteria, so that their results are essentially non-comparable? Or if they study different subjects, with utterly different realms of discourse, such as chemistry and Japanese?

Now there are a number of ways of trying to solve this problem, all of them more or less inadequate. McGaw (1996) summarises them well: use some external examination (either the specific one related to the subject, a single "scholastic ability" test, or some grand total score on all external examinations) to statistically adjust the internal school results; this is statistical moderation of the school-based assessments. Or alternatively "use some external review and checking of schools" assessment results by teachers from other schools or authorised assessment experts to control the level and distribution of school-based results (ie consensus moderation)" (p82).

Such moderation systems provide different processes for modifying the means and standard deviations of school scores on the basis of comparison with other scores or other schools or other students. To the extent that the correlations with the criteria (whether the criteria are scores or actual criteria in the minds of the moderators) are high, to that extent is the moderation reasonable, and possibly invalid. And to the extent that correlations with the criteria are low, or differential, to that extent is error compounded, as we have indicated in the previous discussion.

I do not intend to enter into the debate as to which of these is the "best" way to go, or indeed whether they all do not produce solutions which are more inequitable than the problem they were devised to solve. My project here is not to indicate how such problems may be best solved, but rather to detail what implications such solutions have for the empirical determination of error.

Comparability error

What is clear is that different solutions, including no solution, produce different results. The notion of "true score" is dependent on the notion of some uni-dimensional trait that is obviously non-admissible when the additions involve not only components which have low correlations and do not claim to be about the same thing, but the different additions contain different components. (That is, different additions contain marks from different subjects) But the notion of difference in estimates requires no such theoretical underpinning. It is empirical data demonstrated by differences in empirical rankings or scores under different experimental conditions.

Estimates of comparability errors are easily computed. Given that various forms of inequity are inherent in all measures of both school based and external examinations; that the meaning of the final rank order is based on relative loadings; that all means of trying to create equal loadings involve the creation of arbitrary assumptions and the subsequent construction of additional inequities. Given these facts it is relatively simple to construct a number of different aggregates according to the various models available (including the original raw data), and thus determine the range of ratings (or scores) that these produce. These empirical differences are an estimate of the comparability error. Such a set of scores has the added advantage that it relates to estimates for each individual, and does not confuse such individual differences with group statistics (such as standard error of the estimate).

Note that this is not the assessment error. The comparability error is the additional error added through the procedures of summating or summarising scores, which are independent of other sources of error described elsewhere.

The ontological remainder

My description of comparability error here begs the question as to whether the whole process isn't a nonsense, because of the meaninglessness of the total score. In order to examine that notion briefly I will examine the construct, not of academic merit, which might be a name that we could give to the sum of marks on test or examination performance in various academic subjects, but rather the idea of athletic merit, a similar construct we might conceive in the field of more physico-social endeavour.

> Concerned at the physical flabbiness of our youth, the party in power in the Federal Government, as part of its election platform for 1998, promised to improve the nation's health by removing the flab.
>
> Thus in the year 2000, two lists of year 12 students were produced by Education Departments in each State. One for academic merit, and one for athletic merit. Students are required to nominate three areas of physical prowess. To ensure some breadth they must include at least one area from athletics or swimming, and one from team sports.
>
> Brad and Diana make their choices. Brad, who does not like running, and is not very strong, chose walking as his athletics choice, doubles bowls as his team game, and pistol shooting. Diana chose the hammer throw for athletics, basketball for a team sport, and golf for the third choice. Diana is not very fast or indeed very agile, but she is 1.8 metres tall and weighs 95 kg.
>
> Brad and Diana both covered the curricula designed around their choices, and completed the various tests designed to measure their skills in the designated areas. After some statistical corrections, their separate scores were added to give a final mark. They both obtained the same score of 189 points which is about half a standard deviation above the mean for all year 12 students in Australia.
>
> Independently of this (obviously), they were both offered scholarships at the Australian Institute of Sport; Brad because his pistol shooting scores place him in the world's best ten; Diana because last year she broke the Australian Women's open hammer throw record.

This story is important because it is about individual students and not about groups of students. All of the talk of equal loadings and fairness is in the "equal ends" definition of equity. It attempts to address inequities involving groups of students, but in no way addresses the inequities done to individual students. And just as attempts to address inequities between whole school cohorts invariably leads to other inequities in terms of sub-groups within the school, so any attempts to reduce "better or worse" questions to more or less questions, or any attempt to reduce multi-dimensional entities to uni-dimensional ones, must invariably discriminate against some students more than others, and utterly confuse the meaning of what the final ranking is really about.

The second aspect of the apocryphal story that I want to draw attention to is its obviousness. It is obvious that all of these physical activities are different from each other and that whilst comparisons of aspects within a single sport may sometimes be meaningful, between sports such comparisons are meaningless.

What is not so obvious perhaps is that the complexity and possibilities of difference within cognitive endeavours have much more span, and much more depth, than do those of a largely physical nature. For this field encompasses the whole universe of cultural experience and knowledge. And the ideologies of schooling, if not the practices, assure us that students will have the opportunity to tap this richness. Even so, at the end of the day it all gets reduced to a uni-dimensional list. And both the tragedy and the absurdity of this gets lost in its normality.

Return to Table of Contents

Chapter 11: Rank orders and standards

Synopsis

In this chapter the relationship between rank order and standard is teased out in more detail: In particular the particular meanings given to the standard in the Judge and General frames of reference; how logical confusions proliferate when discourse jumps from one frame to the other; and how the differences in meaning are connected logically.

At the end of the chapter a post-modern myth of the situation is presented.

Personal day-dream

I was about fourteen when I first pondered the sticky issue of the elusive standard. The context was heavenly, rather than earthly, theological rather than educational.

It concerned St Peter. It seemed to me he had a problem. Here he is at the pearly gates as the newly dead file by and do their thing - state their case. And Peter, judge extraordinaire, gives his verdict; pass, fail, pass, fail, fail, fail, etc, etc for millions and millions of people.

And somewhere, among all of those millions were two people, so very close together in the merit of their lives. Oh, so very close! Yet their destiny so very different. For one, just scraping through, the joys of heaven for ever. And for the other, eternal damnation.

But it didn't end there. For as thousands and thousands of years pass, and more and more millions queue at the gate, even between these two he must make finer and finer discriminations.

I didn't doubt he could do it, mind you. Well, it'd be more accurate to say that I considered that if anyone could, he could.

But I wondered why he'd want to!

Fifty years on, these are still the two fundamental questions I have about the notion of a standard : the people who define a standard do in fact have St Peter's god-like omnipotence, but do they have his infallibility? And why do they want to engage in a process that is so manifestly unjust?

Order and standard

Let's go back a bit and tease out this relation between standard and rank order of merit. A relation that I intuited at fourteen, but only recently have systematically thought through.

The relationship is not immediately apparent. There are some judges who are adamant that they can recognise standards and this has nothing to do with relative merit. In fact, to them the word relative is anathema. For them, standards are absolute. They are as solid as a winning post, they are a fact established, a sign as recognisable (to them) as a green light at an intersection. Recognising that some people play games, run races, create rank orders and random distributions and normal curves, they see themselves doing work of a higher order; as maintaining absolute quality in a world trivialised by concepts of the average, the normal, the relative.

So let's push them with a bit of Socratic dialogue. Or is it Hegelian dialectic?

> *You can recognise the standard?*
>
> *Yes.*
>
> *Could you always recognise it?*
>
> *No.*
>
> *So how did you come to reach this state of clear recognition?*
>
> *Through many years of study, reflection, and discourse with other scholars and experts. The senses become refined, the observation sharpened, the criteria established, as slowly, with increasing precision, the standard for quality becomes defined.*
>
> *Let's assume all this is true, and you can in fact recognise the standard. So if I were to show you a work that was well above the standard, you would recognise it as such?*
>
> *Of course.*
>
> *Similarly, if you were to be presented with a work well below the standard?*
>
> *Naturally.*
>
> *It would, of course, be apparent that the first work was better than the second work.*
>
> *True. But this is a consequence of my recognition of the standard, and has nothing to do with its cause. It is, you*

might say, an irrelevant corollary.

Possibly. Now let's take a work that is very close to the standard. You would know whether it was just above or just below, would you not?

Yes, I could make that judgment.

And if I were to present you with another work very close, you would know whether that was just above or just below?

Certainly.

So if one were just above the standard and one were just below, and I were to present you with a third work somewhere between these two, you would know whether is was just above or just below the standard, and you would know that it was between the other two in merit?

I would know that, but only by comparing them all to the standard. Not by comparing them to each other.

Quite so. Now we have talked about five pieces of work. So if I were to present these five pieces of work to you again, you would of course give the same decision regarding each of them.

Certainly.

And incidentally, after the event in your view, you would have them in the same rank order of merit.

Agreed.

Now if they were in a different order of merit the second time, would this not show that there was no absolute standard to which you were able to compare the works?

It would certainly throw doubt on that contention.

And if you can do it with five, in principle you should be able to do it with fifty?

If necessary.

Or even five hundred or five thousand?

Some public examiners do indeed take on that sort of responsibility.

Can we agree then, that regardless of whether the rank

> *order of merit of the works is produced after they have been compared to the standard, or whether the standard is constructed as an artefact of the rank order of merit, in either case the whole notion of standard is in jeopardy unless the rank order of merit is a stable one.*
>
> *This would seem to be a valid argument.*
>
> *Would you be willing to put it to the test then?*
>
> *Put what to the test?*
>
> *Would you be willing to rank fifty pieces of work in their order of merit, (based on their respective distances from the absolute standard) and then do the same task six months later.*
>
> *Me personally?*
>
> *You personally.*
>
> *I'm a very busy person, and it would quite frankly be a waste of time. The result would be obvious. It is self-evident. The orders of merit would be the same.*
>
> *You're certain of that?*
>
> *As certain as I am of my professional competence.*

Now it is apparent that this whole dialogue is in the Judge's frame of reference, and in that frame the notion of an absolute standard logically implies the notion of a stable rank order of merit of all work samples compared to the standard.

It is also clear that the last sentence is not just a rhetorical device, an appropriate metaphor. It is rather a literal truth specified by the very role of Judge. The whole notion of professional competence is dependent on this ability to judge the value of work in the area. To question that competence, then, is to remove the very foundations of the Judge's professional existence. It is an act, therefore, of extreme danger that we would expect to be resisted with great strength, and considerable emotion.

Quality or boundary

In practice our confidence in the standard defined by a Judge cannot be greater than the accuracy with which the Judge can place works, performances, or people in a stable rank order of merit. Our confidence can, of course, be much less than that, but it cannot logically be greater.

That being so, we may think of the standard in two ways: as the lower limit of adequacy, or excellence; or as the line that divides, as the boundary between classifications. Which way we see it is more than a trivial semantic difference. It is an essential point of discrimination between the frames of reference of the Judge and the General, which entail quite different conceptions of the task being undertaken.

For the Judge claims to judge quality, and if necessary the classifications of quality (as inadequate, or good, or outstanding), and the stable orders of merit are a consequence of this.

In the General frame these claims of the Judge are denied. In this frame it is assumed, and the assumption has much empirical evidence to support it, that a judge produces different rank orders of the same works at different times. This indicates at the least considerable fuzziness of standard, and at the most a disintegration of the very concept of the standard. In addition, different judges produce very different rank orders, as well as very different "standards" around which they appear to be, rather randomly and quite widely, distributed. So in the General frame the first task is to stabilise the rank order as much as possible, and then decide the cut-off, the boundary between the classifications of adequate/ inadequate or whatever.

The point that I want to make here is that these two frames of reference are not compatible, and cannot both be used in the same mechanism of assigning a standard without introducing an inherent contradiction into the whole process. The frames are of different logical types; the Judge is a member of the General class. So contradiction is inevitable when the discourse boundaries between them are not clearly separated.

More specifically, we cannot use the General frame of reference to obtain a more stable rank order of merit, and then use the Judges frame of reference to decide the standard, by looking, for example, at some examination papers around what is assumed (from the General frame) to be close to the boundary line. For the use of the General frame has assumed that any judge is inaccurate, and has already produced not a boundary line, but a broad boundary band, within which the Judges' (many and varied and implicit) definitions of standard are to be found.

The price we have paid for the more stable rank order is to make clear the instability and variability of the Judge's "standard." We cannot now go back to the Judge to determine the many (disguised as the few) indeterminate cases by using his/her ability to recognise the absolute standard, an ability already discredited by the assumptions used to make the rank order more stable.

This has not deterred public examining authorities and professional test agencies from doing just that.

Empirical evidence

Facts are less dangerous than theory; despite the promise of the Enlightenment, most people use up far more energy defending their mythologies than in searching for facts; the world is full of answers looking for questions, and significant questions are rather an endangered species.

There is no doubt about the empirical evidence available about the extreme vulnerability of any single Judge in determining either a stable rank order in concurrent rank orderings of the same tests, or in the great differences in rank orderings between different Judges. And this is just for marking. (Hartog, 1936; Cox, 1965; Rechter, 1968; Halpin, 1983)

On the other hand, those plain statements are sanitised by such mathematical constructs as reliability coefficients, some of which become acceptable because they are higher than others; certainly not because they have solved the problem of the stable rank order. In the literature, reliability coefficients of 0.7, and validity correlations of 0.4, are considered very good. They don't look so good when we realise that 0.7 is fifty percent better than chance, and 0.4 is only sixteen percent better than chance.

Now I want to focus on just one aspect of this issue, which relates to the increased stabilisation of rank order obtained through standardised marking procedures, and show how such collusion of Judges produces confusion in the General frame.

The fool-proof marking scheme

The Judge's sense of infallibility in his own ability to recognise standards does not extend to his view of other Judges. It can't, of course, because some of them will disagree with him and then they can't both be infallible. It is necessary then in any particular situation for one Judge to be infallible for all other Judges to be fallible. Thus the requirement in any large scale marking exercise to have fool-proof marking schemes, devised, or at least accepted, by the chief Judge.

In this way the lesser Judges take on some of the aura of perfection of the Chief Judge. And certainly, such schemes do have a considerable effect in stabilising the rank order of students being assessed. And of course, it is easier to determine the detail of such marking schemes in such subjects as Mathematics and Physics than it is in English Expression and Art and History. At least one unused to the cognitive gymnastics of examiners might tend to so believe.

Regardless, a Chief Judge who sets a test paper and then devises a marking scheme could, one would hope, be fairly specific about what content and processes were important, and what criteria were being used to assess the students. These particular values, or prejudices, or idiosyncrasies are then passed on to the other Judges through the marking scheme.

It is obvious that this will decrease the differences between rank orders when papers are marked by different lesser Judges. Statistical data can then be produced showing how "good" marker reliability is. And within the Judges frame it is certainly true that rank order discrepancies have been reduced.

What is not so immediately obvious is that within the General frame the discrepancies have been increased. Within the General frame the rank order shows less variation the more independent Judges there are. The whole point of having many Judges is to "iron out," to balance out, individual discrepancies and prejudices. By effectively reducing the number of independent judges through the marking scheme, the generalizability of the rank order produced to another similar situation is reduced, not increased. For example, we can easily imagine another Chief Judge, with different priorities about the course of study being tested, and different criteria for assessment, producing a very different marking scheme, which would then produce a quite different (though equally consistent) rank order of students.

This problem is not solved, though it may be slightly alleviated, through a more "democratic" production of the marking scheme under the eagle eye of the Chief Judge. The hierarchical structure of the committee, the press to conformity and the expectation of a consensus, will necessarily erode genuine independence on the part of the lesser Judges. Regardless, such "consensus" is not equivalent to the averaging out of independent judgments.

Quantum of error

The Judge can be very specific, at least rhetorically, about what is being assessed. And then the error, as defined by the differences between the rank order produced and that of other independent Judges, is large.

In the General frame, we can reduce the discrepancy between rank orders by averaging out the rank orders produced by a number of independent Judges. But then, because they are individually emphasising different criteria, we cannot be very specific about what we are measuring.

Test agencies and Public Examination systems always assume they are measuring what they are being paid to measure, so regard any improvement in stabilisation of the rank order as a good thing. Persig (1976), in Zen and the Art of Motorcycle Maintenance, assumed that this more "stable" rank produced by averaging was indeed a measure of the elusive "quality" which he sought. I find such interpretations exceedingly suspect, examples of wishful thinking.

The fact is that the more precisely we proscribe one aspect of the intricate web in which the spider variously called achievement or ability or quality of performance lies hidden, the more diffuse other aspects become. We tighten up marking schemes and lose generalizability to other marking schemes. We use many judges and lose specificity about what it is we are measuring. We specify behavioural objectives and lose definition of problem solving. We use multiple choice answers and construction and synthesis gets lost.

We create a test and lose most of what we are trying to test.

This sort of phenomena is well known in the sub-atomic world. According to Heisenberg's Uncertainty Principle, you can know the exact position of a particle, but then you lose information about its momentum. Or you can know its momentum, but then lose information about its position. And the amount of fuzziness, the quantum of error, is a constant. A reason for this is that to collect information about sub-atomic particles, they must be interacted with in some way. And the very process of interaction produces a change in the "original" state.

We are in an analogous situation with tests. The very process of giving a test displaces the person from the "original" situation that the test is meant to describe. We have created an interference by the very process of the experiment, and in so doing have activated an irreducible quantum of doubt concerning our "measures," that can never be appreciated by examining just one measure. On the contrary, reducing the error in just one measure may necessarily increase it in another area. For example, reducing the error in rank order may necessarily increase the error in sampling from all aspects of achievement.

Probably the biggest contribution to this quantum of error is to be found in the boundaries of the test situation itself, regardless of the frame in which it occurs. Such boundaries represent a separation from the everyday learning or working world in which people interact in particular contexts. Knowledge is not something a person has, but rather one aspect of a response, appropriate or not, to a particular environmental context. Test situations invariably remove the person from that real context to produce some sort of controlled, simulated, and hence different context. It is this largely unexamined and unestimated discrepancy that represents a large and irreducible portion in the quantum of doubt.

The enormous popularity (as distinct from reason or purpose) of tests is to be found in its point of congruence with most other myths; in its implicit promise of deliverance from a world permeated with uncertainty, in it's claim to reduce human complexity to a simple story line. In this case the story line of simple numbers.

Judge and jury

> *You haven't really discredited the Judge, you know.*
>
> *I haven't?*
>
> *Of course you haven't. All you've done is to show that some judges aren't as good as they thought they were, and that anyone can be a judge so long as they know something about the topic they're judging on.*

> *So I haven't really got rid of the Judge?*
>
> *Not really. You've just democratised the process of judging. You've let more people into the club, and then asserted that the average of their marks is a better estimate of the true score than the judgment of any one of them.*
>
> *You think I've become a victim of my own ideology?*
>
> *Let me put it this way. If you're convicted of murder, does it matter whether the Judge or the jury convicted you?*

Maybe the metaphor is appropriate. After all, the jury has to make a decision. That is its structural obligation, its very reason for existence. Guilty or not guilty. Those are the choices. So someone, either the Judge or the jury, has to draw the line. After all, they either did it or they didn't. There is a truth to be found. And the Judge or jury's task is to find that truth. Who said that?

The error and the standard

It's at this point that the metaphor becomes shaky. For whilst there was indeed a real crime in the case of the criminal, as evidenced by the dead body of the victim, there is less evidence that there is a real order of merit, a true score. Now if there isn't a true score, then necessarily there can't be a true standard. And even if there is a true score, it doesn't follow that there is necessarily a true standard. As we have seen, the error in the estimate of the standard can't be less than the error is the estimate of the true score. And it will certainly be more, because different judges will differ about where to put it.

> *Ok. So why don't we reduce the error in the standard the same way that we reduced the error in the rank order?*
>
> *How would we do that?*
>
> *Get a number of judges to identify the standard, and then average them out.*
>
> *You mean assume there is a true standard, and then see how well we can estimate it?*
>
> *Isn't that what we did with the rank order?*
>
> *Certainly.*

> *Then why not do the same thing with identifying the standard?*

Now this dialogue worried me a bit when I first wrote it, and it took me a while to ferret out what was wrong with the logic.

Let's start from the beginning. In the General frame of reference, we assume there is a true score, which mirrors a true attainment, or ability, or trait, or predisposition, or whatever And starting from that assumption, we can show, both theoretically and empirically, that we can never measure it. We cannot specify what it is. We can never specify the true rank order of merit. We can only obtain estimates of it, and indicate how far away from our true rank order it probably is.

Now whether there is "really" a true score or a true order of merit of the group being assessed, must forever remain moot. Assumptions of theories do not have to accord with some relationship between variables that have substantive existence in the world. So assumptions of theories related to people do not necessarily relate to any actual qualities or measurable quantities or substantive aspect or observable behaviour of real people. Theories are useful or not according to whether their outcomes, their conclusions, have some links with the observable world. Their assumptions are just that. Assumptions.

However, if we had clear evidence that the assumption was incorrect, then there would seem to be an inbuilt contradiction of our theory to the world that it purports to mirror.

Now if we wish to use the General frame of reference to define the standard, we need to assume that the rank order is the true rank order. For the true standard requires that preliminary assumption.

The claim of the Standard is not the claim of a broad fuzzy space, but of a thin red line. The Standard, if it means anything, means a point on a stable steel scale, not a probability on shifting beach sand.

Defining standards

And we have seen that we can never present the judge or jury with that true rank order. Our own theory had negated the possibility of locating the standard, because it has negated the possibility of finding the true rank order of merit on which the delineation of the standard, in this frame of reference, depends. It is not moot whether the true order of merit had empirical existence. It does not.

> *Well then, it looks as though we're stuck, doesn't it?*
>
> *What do you mean, stuck?*
>
> *We can't use our rank order, inaccurate as it is, to find a standard.*
>
> *Not altogether true. We can define the standard in terms of our true score. In terms of our true rank order.*
>
> *Whose existence is still moot.*
>
> *Exactly.*
>
> *How do we do that?*
>
> *Very simply. If we wish to use grades, for example, we can just define an A as any score or rank order in the first five percent, and an E as the bottom twenty percent, of the population we are testing.*
>
> *Why five and twenty?*
>
> *Make it twenty and five if you like. It doesn't matter. It's arbitrary. The important thing is to define it, so that everyone is talking about the same thing when they're talking about the grade.*
>
> *Won't there be an error in the definition?*
>
> *Not in the definition. The definition is in terms of the true score. So it is exact, as a Standard must be. Of course, in practice there is always an error.*
>
> *So each person is truly at some Standard, but we can never be sure exactly what that Standard is?*
>
> *The second part of your sentence is true. The first part may be true, or false, or just a silly question.*

Reducing absurdity

Let's briefly summarise what we know about standards, and their relationship with assessment, to this point. First of all, we know that empirically an individual judge cannot consistently recognise a standard, nor can he consistently rank students in

BEST COPY AVAILABLE

the same order. These differences between rank orders, and the position of the standard related to them, are increased if different judges are asked to recognise a standard, or rank order students.

The claim of the Judge that he can do these things is thus seen to be untrue as an empirical fact in the real world. It is a fantasy that he has about his own ability that is shared by many people in society. This does not make it less untrue. It does make it less likely that he will admit to its untruth, and more likely that he will take strong measures to disguise the extent of its untruth. For to admit of any error is to destroy the fragile fabric with which the myth of his power and perfection is woven.

In the General frame the error is admitted, though the assumption of an (unattainable) true score is retained. The estimate of the true score is improved by averaging scores from a number of judges. This is vindicated empirically because different estimates obtained by this method are closer together than estimates made by two single judges.

In this frame, it is admitted both theoretically and empirically that any rank order of students is not the true rank order, but an estimated one with built-in error. Thus it makes rational sense to define some standards, some grades, which admit of no error, in terms of percentiles of this true rank order. Even so, in practice we would have to indicate clearly the errors in our estimated grades. And we would have to indicate clearly that these standards are unrelated to any judgments of "quality" as defined by Judges. They are merely cut-off points at various percentiles of a specified population of testees.

What would not be rational would be to get judges to estimate the cutoff points for standards by presenting them with a scale that was admitted to be inaccurate. The Judge claims to recognise the standard, and the production of a stable rank order is a necessary corollary of that claim. We have rejected that claim in our production of a more stable, but still inaccurate, rank order through gereralizability assumptions. It is absurd to now reinstate the judge to determine the standard. It's asking the judge to do something that's demonstrably crazy.

(Not that it's unusual to engage in crazy activities. It would surely be utterly irrational to expect humans to act rationally. The expectation of rationality is the epitome of delusion. It can lead only to despair at the human condition. To applaud rational behaviour in its rare moments of emergence from the mire of human craziness will provide a firmer path to human happiness. But that's another story.)

Judgments and categorisations in the qualitative world

One more point needs to be made here. Whilst the above argument has focussed on tests and grades as a particular sort of educational event, the arguments made are equally cogent for all categorisations of people, whether these be made in the numerical world of quantitative assessment, or in the more linguistic world of qualitative assessment.

Let us be clear about this. If at any point a qualitative assessment engages in a categorisation, a separation of two groups of people, then it is invoking the notion of a standard, and of the measurement of that standard. And in so doing it is logically engaged in all of the rank ordering and judgment errors that have been discussed.

There are some few genuinely dichotomous variables on the basis of which most people may be categorised; for example, blue eyed people and brown eyed people. Most variables used for categorising people however are continuous and not dichotomous; as such, any such categorisation requires a standard, the thin red line that defines the categories, and then a judgment about whether any particular case is above or below that line. As argued earlier, this logically implies a stable rank ordering, which constitutes a primitive form of measurement. Categorisations then involve both standards and measurements, regardless of how much semantic camouflage is used to disguise this.

Democracy and doubt

As the judge topples from his autocratic pedestal of certainty, it is doubtless pleasing to those of democratic mind to know that what will replace the judge is not chaos, but the will of the people. The rule of the individual will be replaced by the judgment of the group. The idiosyncrasy of the individual will be cancelled out and reveal the pure decision of the majority that is the source of the true the right and the just!

We have seen how in practice the delineation of the standard cannot be more specific than the fuzziness of the rank order of those being standardized. And we have seen how individual judges vary considerably in their rank ordering of a group of students, especially if they have no information about them other than the set of examination or test papers. A good punter can (usually) pick a good horse from a bad one, in a general sort of way, but he makes lots of errors when trying to rank accurately all of the runners in a particular race. So it is with the judge of human performance.

There is a crucial difference between the punter and our Judge, however. In the horse race the camera can photograph the finish, so that there is a "true" rank order in which the horses run this particular race. It might not be stable if they run this distance next week, or generalizable to other distances. It will certainly be different over hurdles. But at least in this race we know accurately what the rank order is. Further, we know (almost) exactly what distance they have run, because we have a unit of distance with which we can measure. And we know (almost) exactly what time each horse took to run this distance. If we wanted to, we could nominate a "standard" for this distance below which horses could not compete in the equestrian Olympics. It would be an accurate standard. And it would be arbitrary. And we could measure whether a horse had reached that standard with a small, and empirically determinable, error.

Horse racing as we know it is not a good metaphor for the testing game. So let's develop a better one, a myth more appropriate than that of the infinitely accurate little black box that had mystical knowledge of standards, and resides in the head of the omnipotent judge.

They're racing in Testland

In Testland, races have always been important events. There are no permanent tracks, and unfortunately no way of measuring either distance or time with any accuracy. Some of the more exalted people in Testland do own clocks, but unfortunately they all run at irregular rates, and they all give different times for the same race.

Races are accompanied by due pomp and pageantry. The track is marked with flags and signs saying "this way" and "that way." Horses and riders train hard and are decorated in much colourful finery. There is no starting point and no finishing point but when the bugle sounds they are off and may the best horse and rider win.

There is no actual finishing point, but everyone knows the general area that the race will finish. Here congregate the Judges: the Standard

Judges in their white wigs and purple cloaks impressively flourishing their clocks; and the Placement Judges so serious in their blue serge working suits all constructing their own lines of sight so they can accurately record the order of finishing. Some of these, aware of the subjectivity of human vision, have cameras with which to record the finish in a truly objective way.

In the good old days in Testland there were many more Judges than horses. Everyone would have a great time picking the winner, and recording the orders and times. Then they would happily argue for the rest of the day about who had won and come second and so on. Because all of the judges were viewing the race from different positions and at different angles, because it was unclear which part of the horse had to get past the finishing line to complete the race, and because the signs on the track often had horses running in opposite directions by the time they reached the finishing area, every rider could find some judges who thought they had won the race. So race days were days of celebration and festivity, until . . .

Nobody knows quite when the rot started, when the question about who really won the race became a problem for decision rather than an excuse for argument. Some thought it was when someone suggested that prizes should be given only to the first three horses and not shared equally as was the custom. Others thought it stemmed from a misunderstanding of a remark made by one Sir Henry du Princely, the Queen's sometime lover; another Judge thought Sir Henry said he had the best clock in Wonderland, and took umbrage. But most saw it as the inevitable march of progress and civilisation as Testland lurched forward into an uncertain future; just another example of the dominance of the three e's in the post-industrial era; engineering, efficiency, and expediency.

Regardless of the reason, the facts are clear. Word got around that there was a real winner, and a true rank order in the race. There had to be, because it was self evident that some things were better than others. It followed that some horses and riders were better than others. Thus no-one but an idiot would argue with the blinding clarity of the truth that there was a unique winner, and a verifiable placement order, to every race. The race, everyone knew, was to the swiftest. It became the task of the Judge, therefore, to determine that swiftest.

Sir Henry, who had the ear, as it were, of the Queen, and had been under some flack from other Judges because of the misunderstanding previously referred to, made a unilateral decision that henceforth and from hereon only one clock would be used in adjudging horse races and that one would be his. One or two other Standards Judges who contested this pronouncement found that their clocks mysteriously disappeared, leaving them, clearly, without a tick to tock on, or alternatively a tock to tick on, depending on which University in Testland you went to.

Changes of this magnitude are not implemented easily, of course. At the next race meeting Sir Henry clocked the winning horse and for obvious reasons no other Judge queried his timing. However, the Placement Judges argued that, through no fault of his, he had clocked the wrong horse. Obviously, Sir Henry had underestimated the complexity of the task. He needed the placement Judges in his pocket as well as his clock.

It was at this point that Sir Henry's brilliance shone through with a remarkable insight which ensured his historical survival in the annals of Testland. He let go a double-bunger of a pronouncement that in one foul swoop solved the otherwise irresolvable time and space problems. He defined the finishing line as being where his clock was, and in the direction in which he pointed. By these means Sir Henry succeeded in defining a unique standard and producing a unique placement system at the same time. Truth was now defined. It was what Sir Henry did. He had constructed a new view of reality. A world of winners and losers, scientifically classified.

In conclusion

The astute reader will recognise here the birth of the Judge's frame in its modern form. More importantly, they will see, from their helicopter oversight, that the race has not changed. From above the chaotic nature of the race is evident, and Sir Henry and his little team of supporters can be seen to be doing what they are in fact doing; co-creating a fantasy about a winner where there is none, blinkering vision to substantiate a myth of order, and imposing truth by political assertion.

Return to Table of Contents

Chapter 12: An Inquiry into Quality

Synopsis

From the last two chapters it becomes evident that a fundamental purpose of relating assessment descriptions to standards is to transform notions of quality to notions of quantity. So in this chapter the notion of quality is discussed, and some of the differences with the notion of standard are elucidated.

The theory of logical types is briefly explained in terms of its implications for complex constructs with multidimentional aspects and the special properties of the class "safety standards" is discussed.

The construction of a bridge with various criteria for quality is discussed to illustrate the different languages that must be used to justify the quality characteristics for each criteria. The subsequent history of the bridge is then used to illustrate how the notion of quality is related to boundary conditions and events, and how this affects notions of permanency and attribution.

Some reflections on the nature of quality follow. These are then applied to some of Eisner's ideas about connoisseurship.

Persig's ideas about the metaphysics of quality are briefly discussed, and the relationship between morality and quality on the one hand, and static and dynamic morality, introduced.

All standards are arbitrary

When I was younger and groping for a profession that might suit me, I studied Physics and Engineering. I don't remember much of the detail of those studies, but I did learn two things that are pertinent to this chapter: One is that all measurements contain an error; the other is that all standards are arbitrary.

I remember very clearly struggling with some calculations to determine the cross-sectional area of a steel beam for a bridge. Estimations of maximum loading on the bridge, moments of force and tensile stress resulted in a value of the cross sectional area of the beam accurate to three figures. However, before choosing the appropriate steel T section there was one more step. A safety factor of three must be applied. Or was it four? No matter, the calculated cross-sectional area must be multiplied by this arbitrary number in consideration of possible tornadoes, earthquakes, rock concerts on the bridge, or whatever other natural disasters might inadvertently occur. This undoubtedly would make the bridge safer for traffic and incidentally more profitable for the steel manufacturers. And it made the accuracy of the initial calculation absurd.

Safety and quality

At this point I want to try and untangle another confusion that has bedevilled the notion of standard, especially as applied in the human sciences. This is the confusion between safety standards and quality standards.

In the manufacturing area there is less confusion. Standards that apply to car seat belts, bumper bars, brakes, lights, are clearly basic safety requirements. General design of car, colour, control panel layout, type of upholstery, fuel economy, are aspects of quality. And of course, one aspect of quality is that all safety standards are met.

Safety is about prevention. Safety is about what is not, about events that are always immanent, yet, if safety is successful, never materialise. Safety is about the future that is frustrated, about unrealised potential. Because each safety measure blocks a road to disaster, each safety measure is essential in its own right. To meet a safety standard is to claim that one such roadblock is in place. To know that all such safety standards are met is to be reassured and insured against disaster. However, to know that eighty percent of safety standards are met is to know nothing about which particular safety standards are not met. For a gambling man this may be a situation of high desirability, and hence provide an experience of high quality. But in the world of safety standards, this is a recipe for disaster.

Quality on the other hand is about manifestation, about potential realised. Quality is not so much about specific aspects as about their interrelations; about interpretation rather than measurement; about the whole gestalt rather than summaries. Further, notions of quality are intimately and necessarily connected with the observer, and hence are constructed from the observer-object interaction, rather than claiming to be a measurable component, or sometimes a presence or absence, of the object or specific attribute being observed.

Theory of logical types

The theory of logical types is about levels of abstraction in human discourse. One of its axioms is that whatever involves all of a collection must not be one of the collection; that is, that there is a fundamental distinction between a class, and the members of that class. This might seem obvious. Obviously a single man is not all men, and a married woman is not all women.

Trivial as this might seem, the conclusion from the theory is far from trivial: that when this clear separation between class and members is not made, messages become confused. As Bateson (1972) describes it, "the theory asserts that if these simple rules of formal discourse are contravened, paradox will be generated and the discourse vitiated" (p280).

Human discourse is decidedly more complex that simple logical syllogisms. We do not usually talk like logic machines. We talk very often in and about abstractions, and these abstractions may be at different levels of logical type. We present information (first level), and give an interpretation of that information (second level), in a particular context which affects its meaning (third level). A story that makes fun of a rich Jew has a very different meaning if told by a speaker at an anti-semitic rally than it does when told by a Jewish comedian on a New York stage.

Of particular interest here is that errors that lead to confusion occur when the properties of a class are ascribed to members of that class, or vice versa; or more subtly, whenever the discontinuity between class and member is neglected, and they are treated as if they were at the same level of abstraction:

> The theory of Logical Types makes it clear that we must not talk about the class in the language appropriate for its members. This would be an error in logical typing and would lead to the very perplexing impasses of logical paradox. Such errors of typing can occur in two ways: either by incorrectly ascribing a particular property to the class instead of to its member (or vice versa), or by neglecting the paramount distinction between class and member and by treating the two as if they were of the same level of abstraction (Watzlawich, 1974, p27).

Safety and logical type

Safety is not quality. It is one criteria we might use in describing quality. It is a member of the class of such criteria. But it is a very particular member, because it is atomic in its construction. It is comprised of a number of specific safety requirements each of which must be individually met. Not only is the class of events or information called "safety" of a different logical type to the class called "quality," but the essential information about safety is lost when the class "safety" is described, rather than the individual items that describe it. Unless, as we mentioned earlier, the statement about the class is that "all safety measures have been satisfied."

Safety and people

In many aspects of our life safety measures are important for its continuance. In home, leisure activities and job, safety requirements contribute to our health and that of others. So matters of safety are a part of various educational programs. As such, it would seem important that evidence be obtained that students have incorporated such safety items into their behaviour. Or, at the very least, that they understand and can implement all of the safety requirements. Talk of safety (like talk of sexuality) produces points of high density in the field of power relations.

It should be apparent, however, that test or examination information involving rank orders or grades or marks regarding safety represents information about the class of safety items, and as such is inappropriate and confusing. If safety requirements are essential requirements, then marks of 70 per cent or grades of C for safety, or for tests which include questions about safety, present information that is inherently contradictory. By definition, if you have not met all safety requirements you are unsafe.

Test-makers and others argue that in the context of a test people make errors and it is not reasonable, because it rarely happens, to expect one hundred percent correct response. This is surely an indication that the test context is inappropriate for obtaining information about a person's acquisition of safety measures. It certainly does not justify accepting that if they can provide evidence that they "know" seventy percent of the safety requirements that their "standard" of safety is adequate.

Further, talking about safety measures, or choosing the correct safety requirement from a number of choices, is an activity of different logical type than implementing that information in the context of a job. Talking about something you do is of a different logical type than doing it. So any measure on a test, even at one hundred percent, cannot be a measure of safety behaviour. It is a measure of test behaviour. At the very best it is an indicator, about which empirical evidence could be obtained about the probability of its correspondence with overt safety behaviour under specified conditions. In this respect, probabilities less than one would necessarily indicate test invalidity.

Safety and minimal outcomes

The idea of minimal outcomes is analogous to that of safety. Minimal, or minimum, means the least amount, the lowest possible. If a course of study has a set of minimal outcomes that define its successful completion, then by definition all such outcomes must be demonstrated if the course is to be satisfactorily completed. To set a test incorporating questions related to such outcomes and then use a test score (a statement about the class) to describe the "standard" that has already been described by each of the members of the class, is again to confuse logical types. Such tests are sometimes referred to as mastery tests.

There are three additional confusions, two of them the same as for "safety." The first is that only a perfect score is consistent with the definition of minimal. So to attempt to find an appropriate "cut-off" score to use as a standard is to engage in a paradox, is to indulge a contradiction, is to professionalise an absurdity. Berk (1986) was able to identify 38 methods for setting standards and produced a consumer's guide (to choose the most appropriate absurdity).

The second confusion involves the fact that context affects meaning. For many educational outcomes the context of a test situation is inappropriate anyway and represents another logical type confusion. For example, any outcomes involving verbal discourse, such as listening skills, group problem solving, giving instructions, cannot be demonstrated in a written or multiple-choice test without logical type confusion occurring. Writing about verbal interaction is not verbal interaction. Choosing the most appropriate response from a multiple-choice selection is not responding oneself in an interpersonal context. Talking about a painting is not painting. The whole test and examination industry is permeated with this sort of confusion.

The third confusion is one of ends and means, and is well described by Burton (1978): "no measure of a single skill can ever be mapped on a non-trivial vision of real success because any problem can be solved in more than one way. One can determine whether the respondent has the skills necessary to solve the problem this way, but one lacks the justification for imposing successful performance, this way, as a standard"(p273). Burton believes that "this argument is fatal to any method of setting performance standards." Burton is perhaps mistaken in believing the issue is amenable to rational argument, and does not consider that it may be entrenched in mythical discourse.

Mastery tests and frames

Mastery tests result in scores produced by the summation into a numerical score of specific objectives attained. In relation to error, they contain all of the errors of specific objectives plus a large labelling error. In adding the results most of the important information is lost, in that we no longer know which specific objectives have been attained and which have not.

In this situation, whilst the generation of the test has used the Specific frame of reference, the summation has resulted in a normative test score. We no longer have information about what a student has achieved. We have information only about how many of the objectives have been achieved. This is exactly equivalent to information about how many addition sums are correct, or how many words are correctly spelt, or how many formulas in dynamics we can remember. The description is now clearly normative, and may only be interpreted in terms of whether one student got more or less "right" than another, or in terms of some arbitrary "standard" of how many "correct" answers will be considered "adequate"; how many correct answers constitutes a "pass."

In this situation, because information about the particularity of objectives attained is lost, the whole detailed descriptions tend to be similarly "lost," or unavailable to those interpreting the test information. Labelling errors thus become large, as the meaning of the score, and the label attached to it, are differentially interpreted.

Mastery tests and internal logic

In most courses there are some facts, some understandings, some activities or skills, which are central to what the course is about, so that we could say - if they don't know at least those things, or if they can't do at least these things, then there is no way we could say they have adequately completed the course. In old-fashioned terms, they are the "must knows" or "must dos" of the course. As distinct from the "should know" or "could know" categories.

Now there may be some areas of study where curriculum writers or teachers are unable, or unwilling, to specify such a category of "must know" performance. However, when it is so specified, it comprises a description of a finite number of procedures or products that will demonstrate the "knowing" of these crucial things. In other words, within this limited "must know" area, it is possible to specify what must be done, the conditions under which it must be done, and the procedure by which its adequacy will be known.

These then could be used to describe the essential requirements of the course of study. They are limited in number and extent, and are specifiable in the specific frame of reference. As they are accomplished, as evidence is obtained that each outcome has been achieved, this can be certified by the teacher or student. If there are ten such outcomes, then successful completion of the course would require that all ten outcomes be so certified. Otherwise they cannot, obviously, be essential. To certify that eight out of the ten essential requirements have been completed is to certify that two of the essential requirements of the course have not been completed, and thus to certify that the student is uncertifiable. More than this, it is to lose the information about which two essential requirements have not been demonstrated.

So to obtain a "total score" on a mastery "test" is to contradict the whole concept of essential requirements, and to lose all the relevant information. Unless the total score is a "perfect" score.

In many situations the very notion of a "test," of some particular situation constructed to check all of the essential requirements at one time, would itself be contrary to this frame of reference. In the artificial and often pressured "test" situation it might be expected that success in some "essential" activities might not be demonstrated. It is this very argument which has been used to justify the acceptance of less than a "perfect" score in a mastery test. Rather it should be seen for what it is - an argument that invalidates the use of the test.

The problem of time-binding is not solved by success in test situations any more that it is by success in the ongoing teaching - learning context. We can never certify that any fact will be recalled at a later date, that any understanding will be retained in the future, that any skill will be demonstrated again successfully next year. We can sensibly certify that a behaviour has occurred once, or twice, or if necessary one hundred times. Regardless, we can never be certain it will be adequately demonstrated on the next occasion.

Test givers imply, with their insistence on testing, that demonstrations outside the testing situation are in some way of limited value, credibility and validity. It has always seemed to me that "tests" have all the inadequacies of "on site" or ongoing certification, with quite a few bonus inadequacies added on for good measure.

Or more accurately, for worse measure.

A bridge of quality

Let's assume that we want to describe a particular person's performance in a certain area. Building bridges is as good an area as any. And we are interested in the quality of that performance. That is, we are in the area of discourse often called assessment.

We might decide that there are four aspects of performance which we want information about; four members of the class we will call quality; four criteria on the basis of which we will assess quality of the bridge produced. Is the bridge safe? Is it economical in cost of materials, construction, and maintenance? What is its environmental impact in its rural context? And how is its aesthetic design judged in a competitive order of merit in relation to other submitted designs?

We note in passing that this decision about these particular four aspects of quality is itself a value judgment subject to enormous error in the General frame of reference.

It is clear that the language of discourse for each of these four criteria will be different, and attempts to simplify by means of some language that is appropriate to some and not others, or that is appropriate to the notion of "quality" as a class but not to some or all of the members of that class, is to compound confusion by oversimplification (Eisner, 1991, p182).

For example, the first question, about safety, may only be addressed by showing that all safety measures are in place; the language that designates individual safety standards is appropriate. The question about being economical involves careful costing; the language of accounting is appropriate, and the language of economics will be necessary to delineate boundaries. The question about environmental impact will draw information from a number of disciplines - geology, biology, ecology, geography, ethics, economics, and so on. Ultimately, the discourse must deal with the balances and trade-offs among conflicting values and pressures; the language of politics and the language of environmental ethics will fight it out. Finally, the order of merit based on the aesthetics of the design will draw on the language of art and architecture, and be involved with issues of the assessors' personal tastes and the profession's current fashions. Finally, however, such complexities will be reduced to a single dimension where better-worse becomes more-less and a rank order is produced.

As this competitive order of merit is one aspect of the quality of the design, it is not that quality. By the same token, no measure of the order of merit can be the measure of quality, any more than a cut-off point on the order of merit can represent a cut-off point of quality. All this regardless of how consistent, stable, generalisable that order of merit may, or may not, be.

Permanence of quality

> *I've been thinking about the quality of the bridge.*
>
> *The one where I chose four rather arbitrary aspects of quality to talk about?*
>
> *Yeah, that one. You made it easy for yourself by choosing something very practical and material and solid. I mean, it's stable, you can see it and jump on it. It'll still be there tomorrow so that others can assess its quality for themselves.*
>
> *It does have that illusory aspect of permanence.*
>
> *Why illusory? A bridge is a pretty permanent structure.*
>
> *Even so, the notion of quality is somewhat ephemeral. Let's see how our bridge, built five years ago, has stood up to our quality assessment. First the aesthetic quality, the only one subjected to the rigours of competition, of rank ordering and the notion of the standard. The design was brilliant and quite spectacular. There was some controversy after it was built about its enormity. But mostly there was approval. Then, of course, fashions change. Most "experts" these days consider simplicity a major design virtue.*
>
> *You're saying that if the competition were rerun today this design wouldn't have won?*

That's what I'm saying. These days big high ornate bridges are out. Simple low bridges are in.

What about environmental impact?

There's the bridge's visual domination of the landscape, which is much more intrusive than was anticipated. The terrain is very flat. So you can see it twenty kilometres away. But more important for some is the impact it's had on the lesser crested poorigal. The bridge has affected its navigational ability in some mysterious magnetic way. Apparently this area was significant to a change in direction during their yearly migration. Now they fly in circles around the bridge till they drop. Suddenly they've become an endangered species.

What about the economic question?

Interest rates have gone up by a factor of three, they've put a toll on the bridge, and the government has had to bail out the Roads Board once already. What was once an economic asset has become an money-eating monster.

Well, I guess fashion, the environment, and the economy are always a bit suspect in terms of their stability. But at least the bridge is still there, and it's safe.

Not exactly.

What do you mean, not exactly?

Just one of those unfortunate things really. It's not considered a major earthquake area. Almost no activity over the last sixty years. Then last week there was this major tremor. Point eight on the Richter scale. A major fault line developed just a kilometre away from the bridge.

Did it damage the bridge?

Not exactly. Amazing structure really. Shows how good the design was. Not a crack anywhere. Only one problem.

What's that?

When the land tilted, the whole bridge tilted with it. The road slopes thirty degrees.

So what happens now?

Well, the bridge is useless. The only question now is whether to leave it there, or spend half a million to blow it up and

> *remove it, thus saving from extinction the lesser crested poorigal.*

The apocryphal nature of this story does not diminish the fact that the bridge, like everything else which has a material presence on this planet, is not permanent. It will change. It is not fixed in space and time. The rate at which it is ravaged by time - that is, by the events that indicate its interactions with the environment - is normally quite slow, and hence our sense of its relative permanence compared to our own brief life-span. Yet in geological times the life of the bridge, as a bridge, is minuscule.

What is important to understand about this very sad story is that it indicates very clearly that the bridge itself does not have any qualities. Putting it another way, none of the qualities we discussed in relation to the bridge belong to the bridge. They are rather descriptions of how the bridge will interact with other things - with the physical and geological environment, with the economic system utilised to finance it, with the human cultural world in which it is enmeshed. So when any of these environments change from those expected, so does the quality of the bridge.

Nor does the bridge have some aesthetic qualities having a magical existence independent of the bridge and its environment. You may conceive the bridge as being beautiful, as some music that you hear is beautiful, or the second law of thermodynamics seems beautiful. And indeed there may be a palpable human response that you have to these three events which justify using a single word, beauty, to describe them. Even so, it is clear that the similarity is contained in your particular response to the events, rather than to the objects that are responded to.

All of which does not mean that beauty is in the eye of the beholder. To take that view is to denigrate the object. Just as to ascribe the beauty to the object observed is to denigrate the observer. If the label of beauty is to be pinned anywhere, then it must be pinned to the event, the interaction, the relation, between observer and observed. Qualities, like any other form of data, are constructed from events, not discovered in objects.

Quality, standard and logical type error

Let's look then at what might represent quality in a teacher or student in a school.

The function of the school is not only to prohibit and punish and exclude but to produce. To produce good work. Though even here, good work is but a symptom of the more important school product, the good student. The good individual student. Increasingly, it is not so much the work of the student that is valued, but the "whole person" that presages it. Abilities, attitudes, skills, the whole plethora of attributes fantasised to define the good student, the good worker, the good manager, become the focus of attention, the point of application of the standard.

This is not new, though it is more overt that it was twenty years ago. I remember doing some consultancy work in a Primary Teachers College in the 1960s. I visited the various faculties, and talked to the lecturers. Indeed, they were concerned that the students had sufficient knowledge to teach the subject. But what was more important was that they had a very positive attitude to the subject, that they really liked teaching mathematics, or music, or history, or science, or physical education, or whatever. On the surface, a useful intent. Yet when I tried to picture what sort of a person this would be, with great enthusiasms for everything that they taught, I could see a successful sales-person, but hardly a successful teacher.

It was laudable that these lecturers communicate their enthusiasm to their students. It was their inability to see its overall implications, and its curtailment of any critical thinking on the part of the students (or indeed often on their own part) that was cause for concern. My problem was to discern the difference between a student enthusiastic about the whole curriculum, and a happily conforming blob.

The error is a logical type error. In the class "quality" there are many members; there are many aspects of a person that relate to quality performance. One of these may relate to the particular context. Another may relate to standards of proficiency. Another to integrity of values. The language of discourse of these three areas will be different. But all of these discourses must be both utilised and transcended in a discourse on quality, and no measures of the members of the class (assuming such measures are possible), can be a measure of quality.

Another example; quality of life is not the same as standard of living; there is a world of difference, indeed a life-style of difference, in the two concepts. For the very essence of quality is its immeasurability, its identification with a world not wholly material, an association with that mysterious realm of experience called "soul." Quality is concerned both with essence, with experience from within, as well as with experience perceived through reflection from surfaces. Standard of living, on the other hand is a function of measurable quantities; income, savings, washing machines, televisions, supermarket shopping bills, and whatever; the countables, the quantifiables, of the material and materialistic world. Again, "standard" is a member of the class "quality." And for that very reason the two concepts cannot logically, and hence rationally, be identified.

Adequacy and labelling

How do we solve the dilemma? If standards cannot do the job expected of them, what do we replace them with? The issue of competence in a job does not go away because of the errors and confusions in its measurement. On the other hand, it is possible within a particular milieu for a group of people to agree with some consistency, and hence certify, that certain work has been carried out adequately. In every family, in every school, in every sporting team, in every job, work is done and considered adequate. It is useful for some purpose and not dangerous. And the conditions of that work, (and hence of that agreement), may be democratic or elitist, may press towards convergence or divergence. In other words, there is a notion of adequacy, or competence, or comparative excellence - in short, of a limited sort of quality, that is both embedded within and produced by any work culture, in terms of which individual performance is assessed. What is also clear is that this notion is fuzzy and multi-dimensional, error prone, describable rather than measurable.

What becomes clear here is that this notion of adequacy, of quality of the work, is not independent of the culture in which it occurs. The label of adequacy is a label belonging to the whole interactional milieu in which the work occurs; yet another reason for the immense errors that become apparent when such work performances, or the abilities or skills or predispositions or aptitudes that are fantasised to explain them, are pinned onto particular workers, and to a lesser extent on particular criteria or products (Fielding,1988; Raven,1992).

Quality

> Quality . . . you know what it is, yet you don't know

> what it is. But that's self-contradictory. but some things are better than others, that is, they have more quality. But when you try to say what quality is, apart from the things that have it, it all goes poof! there's nothing to talk about (Persig, 1976).

Maybe the apprehension of quality really is a mystical experience. And maybe not. On the basis of the discussion so far, I will try to give the skeleton a bit more flesh.

Quality refers to a particular experience. The notion of quality is a complex one, involving a number of aspects of the experiential event that can be discriminated. The possible aspects that could be discriminated always exceeds the actual aspects discriminated; an informed choice is made about what particular aspects will be discriminated in this particular case. The choice itself is arbitrary, in that different choices could have been made, some of which would in retrospect be approved. Such choice of course mirrors value.

Discourse about any one aspect might or might not refer to some standard of accuracy or adequacy or competency or whatever.

Balance or harmony or elegance is an aspect of quality. This involves the relationship between the aspects initially discriminated. All this so far is a description of surfaces, of what the object or performance appears to be from the outside.

How does this relational aspect look from the inside? If quality is more the spirit of the product (the person, the event), then quality relates to the interior of the holon. Quality is, in human terms, the expression of the life force immanent in the product, or in the production, or in the person in the process of production; that is, in the production event. Quality then becomes related to a state of consciousness, or its analogue in non-conscious productions. It involves the integrity, the meaning, both of the producer and the product.

Quality also involves the integration of the inside and outside; the aligning of truthfulness with truth; of inside and outside awareness; of the aligning of the potential of the stone with the vision and skill of the sculptor; of the sound of the spirit with the song of the singer (Wilbur, 1996).

From the inside quality is experienced as the essence of the event, of the spirit of the relational experience. It is thus the meaning of the event as interpreted by its participants. It may be, indeed will be, different to other similar eventful experiences, and because of its idiosyncrasies is not comparable to them in any linear way. So it is not possible to link this notion of quality to ideas of adequacy or competence or of other categorisations which necessarily involve standards. What words then are suitable? Beauty perhaps? Elegance? Flow? Life? Spirited? Words that describe the essence of the experience, of the connection!

In relation to people's performances, the notion of quality can be attached either to the creative process of the performance, or to a particular product of the performance. Post-structural analysts want only to attend to the latter, regarding the former as irrelevant. And of course the event that involves a critic interacting with the product is a different event to that event which produced the product. As such the qualities of the two events are necessarily different and essentially non-comparable. The element they have in common is the final product; but this product was the culmination of the first event; it did not exist till the final moment of the first event.

On the other hand, it is sometimes a stable and reproducible element of the second event. The two events are holarchicaly connected. The first event (culminating in the product) can exist without the second (the critique). But the second event cannot happen without the first. It follows, as with all such holarchical connections, that the attributes that determine quality in the first event are not necessarily or probably those which determine quality in the second. They are different creative endeavours; they have different intentions and languages; to misrepresent this difference is to court confusion.

Eisner, quality, judgment and standard

Eisner is one of the few writers in the assessment field who has attempted to analyse in depth the notion of quality through his notion of connoisseurship. Eisner (1991) differentiates qualities from qualitative from quality. "By qualities I mean those features of our environment that can be experienced through any of our senses"(p17). So a quality pertaining to a person is any aspect of that person on the basis of which we can differentiate by using our senses. "Aspect" or "attribute" or "property" may be better words to use because they avoid the confusion with the notion of quality we have been discussing. He goes on to claim that "we can only appraise and interpret what we have been able to experience," but then warns that "if our perceptual experience is aborted for the sake of classification, our experience is attenuated"(p17). Eisner adds that "the qualitative aspects of experience are not only secured in attending to qualities out there, but also are manifest in the things we do and make"(p18). In my terminology, aspects are discriminated both in the event that produces a product, and in the event in which it is perceived.

"The ability to make fine-grained discriminations among complex and subtle qualities" is what Eisner (1991, p63) calls connoisseurship, the art of appreciation. The art of recognising quality, as I am using the term. He recognises a fundamental problem with his notion of connoisseurship:

> we may find critics with very different views of the same situation or the same book. What are we to do with such differences? In standard research methodology, we might dismiss the critics as incompetent and find new ones who can independently agree, or we might look to our own criteria and methods, for these might be at fault. Our methods might not be clear or, if clear, they might be incomplete, or our instructions to our critics (or judges) might be ambiguous. The point is, we would not trust differences of view; such a circumstance indicates statistical unreliability. We would try to achieve reliability among judges. As a last resort, perhaps, we might decide to limit what the critics were to attend to. By simplification we might achieve a higher level of intercritic agreement, even if in the process we compromised validity (p113).

Obviously, Eisner does not agree with this response, and is critical of it. "Critics might be attending to different dimensions of the same work," he points out. They might be bringing different perspectives to it, be sensitive to different aspects of it. No one knowledgeable in literature, "would dream of trying to calculate a mean among critics as an adequate test of a critic's work"(p113). Maybe not, but such consensus is often seen as an adequate test of the work being criticised, and that is the issue here.

And indeed, that is Eisner's test for the adequacy of the critic's work: "consensual validation in criticism is typically a consensus won from readers who are persuaded by what the critic had to say, not by consensus among several critics"(p113).

What is such local consensus except a qualitative calculation of the mean? And note how the second order consensus has distracted attention from the first order contradiction, to which he does not return.

Why are collections holding contradictory judgments so difficult for Eisner? In his criticism of specific behavioural objectives, Eisner (1985) says that those who evaluate them "often fail to distinguish between the application of a standard and the making of a judgment" (p115). He then quotes Dewey, who, he says, "makes the distinction quite clear." So what is the distinction according to Dewey? Standards, according to Dewey, define things with respect to quantity. And measuring a quantity is not itself a mode of judgment.

And qualities are qualities of individual objects, even though the critic reveals himself in the criticism. So to Dewey, and Eisner, the qualities are indeed inherent in the individual object, even though the description of those qualities is enlightened by the connoisseur. And nowhere, concludes Dewey, "are comparisons so odious as in fine art" (Eisner, 1985, p115).

So Eisner is clear that qualities cannot by measured by standards. And of course they can't, because standards are definitions and not measurements. What he must mean is that qualities cannot be measured by comparing with standards, both because measurements and judgments are of a different order, and because comparisons are odious.

So he is trapped; qualities are inherent in the object; connoisseurs make the fine discriminations that enable them to describe quality; such judgments are not measurements and abhor standards; even so the judgments might lead to categorisations of the object (of winner of the contest, or worth a distinction, or inadequate at this level), which bypass standards and measurement. Yet connoisseurs differ sometimes fundamentally in their categorisations.

I have argued in the previous chapter that such categorisations necessarily invoke standards, and comparisons with them. But even if they don't, two contrary judgments of connoisseurs create a contradiction that denies that connoisseurs can categorise accurately, and this is surely one of the essential aspects of their connoisseurship. An alternative explanation, of course, is that the qualities do not reside in the object, but are rather an aspect of the event that involves the interaction of the object with the critic. In which case to categorise the object is to mislabel the event, and hence by implication to mislabel the person who produced the object.

All of which takes us back to Eisner's original question: What do we do with such differences? Eisner says don't do what is usually done. And then is silent. Maybe if you ignore them they'll go away! I note that he is talking about consensual validation in this section of the book, and validation, as we have seen, is an advocacy argument for the defence. It follows that the disagreement has to be ignored, because it represents the essence of the (unspeakable) case for the prosecution (See Chapter 16 on Validity).

Summaries or collections - the crucial choice

So Eisner doesn't want to celebrate difference as being at the cutting edge of new knowledge, the collection being the best description, superior not only to a summary, but also to any consensual agreement. For to do this is to deny the possibility of the accurate categorisation of people or their creative products. And that is the cutting edge of the power of the connoisseur. Such power does not ultimately lie in the cogency and plausibility and depth and sensitivity of his critique, however much the connoisseur may wish to believe it is so, and even though this advocacy may well support such power; in practice it lies in judgments that define the standards that produce the categorisations that determine the lives of Jack and Jill and all their little children.

This necessity to categorise in a single dimension is illustrated by Rosenberg (1967). In his book On quality in art, he looks at criteria of excellence from the 16th to the 20th century. He quotes de Piles, a 17th century critic, who:

> evaluates the best-known artists of the past and present in a very special way: the artists are graded in each of four categories already mentioned (composition, drawing, colour, and expression). He scores each category against an ultimate grade of 20, which would indicate perfection (p36).

He then goes on to say "de Piles does not give us the sum total for each artist." Presumably it never occurred to him to do so. But then Rosenberg adds: "but we can easily do the addition"(p36). Presumably, as a child of the 20th century, it never occurred to him not to.

Rosenberg (1967) then uses this magical and meaningless sum total to criticise some of de Piles' ratings; "We are disappointed that he rates Michelangelo (37) much lower than Andrea del Sarto (45) . . . We cannot understand why Durer receives a grade of only 36, when a second rate Mannerist like Taddeo Zuccaro gets a total of 46"(p37). And so on. But of course de Piles gave no such grades. He knew it was meaningless to add a mark for colour to a mark for composition to a mark for drawing.

In assessment, whether qualitative or quantitative, the crucial choice made is whether to opt for summaries or summations on the one hand, or for collections on the other: to opt for summaries is to go the way of simplicity, of communality, of "truth." A summary celebrates similarities by defocussing differences; to opt for collections is to stay with complexity, with uniqueness, with essential uncertainty. A collection celebrates differences by defocussing similarities.

Summaries and summations then are basically conservative; they are uni-dimensional; they are dedicated to notions of order and security. Collections are basically radical; they are multi-dimensional; they are dedicated to notions of creativity and anarchy (in its positive persona).

To date, the history of educational assessment has been a developmental history of the summary. The current agony of many of its most thoughtful protagonists (Delandshere, 1994) will only cease when they settle for collections, and deal openly and ethically with the personal and social consequences of that choice.

Assessment of quality as moral action

Persig (1991) makes a strong link between morality and quality; in fact, to him they are synonymous terms.

He looks at the relationship between evolutionary structure and the metaphysics of quality, and shows that there is not just one moral system, there are many: In the metaphysics of quality there's the morality called the "laws of nature," by which inorganic patterns triumph over chaos; there is a morality called the "law of the jungle" where biology triumphs over the inorganic forces of starvation and death; there's a morality where social patterns triumph over biology, "the law"; and there is intellectual morality, which is still struggling in its attempts to control society.

Each of these sets of moral codes is no more related to the other than this dissertation is to the flip-flop circuitry which controls the computer on which it is typed. Let's consider this in relation to our bridge; its quality as a physical structure in the inorganic world was unrelated to its quality as part of the social life of people; just as that in turn was unrelated to its quality in that intellectual world that can conceptualise its probable long term effects on the environment, and hence on the lives of humans not yet living.

Further, there will often be conflicts between the static social morality that would hold the physical or biological or social structure stable, and the dynamic evolutionary morality that would move it onward:

> Intellect is going its own way, and in so doing is at war with society, seeking to subjugate society, to put society under lock and key. An evolutionary morality says it is moral for intellect to do so, but it contains a warning; just as a society that weakens its people's physical health endangers its own stability, so does an intellectual pattern that weakens and destroys the health of its social base also endanger its own stability(Persig ,1991, p168).

In a morality based on stasis there is no confusion; what destabilises the social system is immoral, is an act of inferior quality. Yet in a static-dynamic view of evolution this equation no longer holds. The central problem then becomes, in Persig's (1991) words:

> How do you tell the saviours from the degenerates? Particularly when they look alike, talk alike and break all the rules alike? Freedoms that save the saviours also save the degenerates and allow them to tear the whole society apart. But restrictions that stop the degenerates also stop the creative Dynamic forces of evolution (p228).

It would be easy to say that the actors themselves are aware of whether they are saviours or degenerates, but this is problematic. There may be cases of genuine manipulation, of intentional evil, but these are probably rare. Most choices are internally processed as the competition of two positives, not as the balance of good against evil. And even when the latter is the basis of the internal dialogue, the "evil" may often be a societally imposed value that from another frame of reference could be seen as positive.

In both cases, the actor must act on a sense of "rightness," of "necessity" that overrides choice. The actor, like the observer, simply cannot tell what the ultimate quality of the action will be, because the actor can never predict all the consequences of action. To claim that the ultimate test is whether the act is free of ego is to beg the question. Any act can be interpreted as ego-dominated, even acts of transcending the ego, which are designed to nourish the "super - ego."

Finally, we are left alone with our own sense of identity, our own sense of integrity. After all the agonising, all the reflection, we are finally left with a sense of the flow of life, with the flow of one particular life, of one particular relationship; with a sense of appropriateness that on the basis of static moralities is sometimes most inappropriate. And we do what we must do. This is the essence of evolutionary morality; it is the essence of what constitutes quality in the intellectual sphere; it is the essence of the meaning of quality in any assessment event in which a product or a person is the focussed element. It is a demonstration of what Churchman (1971) and Campbell (1956) call the heroic mood.

Quality products

Traditionally the problem of the relationship between quality and standard has been solved either by ignoring it, or by emersing it in semantic confusion: by fuzzing the boundaries, by assuming the two concepts are isometric, by ignoring the logical type error, by claiming that high standards are of course synonymous with high quality. And as it is self evident (within mythical discourse) that we can measure standards, it follows that we have measured quality.

What we have done is something much more damaging; by identifying standard with quality we have confined quality to the straight and narrow, and thus denied its very essence, which is to be found in its spontaneous deviation from the constraints of geometric efficiency. For the standard is a preconceived point (however practically unmeasurable) on a predetermined scale. It may indeed be used to describe a work of conforming excellence, but is quite incapable of recognising the nuances of diversity, the force of spirit that transforms articulate parrots into creative people. One of the characteristics of works of high quality resides in their difference, not of measure, but of style. Quality is perceived not in differences in kind, but its differences in difference; not in differences in length, but in variations of depth: in short, quality diverts us from the linear, takes us to a dimension orthogonal to the flat. "Quality, consciousness, and experience are separate words for what is one whole, as one lived-process" (Beittel, 1984, p110).

The essence of quality resides not so much in the aspects or characteristics with which we attempt to describe it, but rather with the relationships between those aspects, and the coherence of the whole gestalt that those relationships produce, and hence with the meanings that such coherence implicitly evokes. And as with all gestalts, it is recognised as such only within the milieu of its production, only against the culture that is its backdrop, only in terms of the event through which it emerges. As no two products in this material world can ever be completely identical, so must the quality that characterises them also differ. As that quality is multidimensional, and contains relational aspects, it is idiosyncratic to each product, as well as to the conditions of its production.

In general, discourse on quality is not amenable to that "better and worse," "more or less" description that is a prerequisite for any measure, and hence of any standard, or any categorisation. It is sometimes amenable to discourse, and to aesthetic response, and even to comparison in some of its aspects. And quality is amenable to change, both in its own meaning, and to the meaning it generates in relation to the product it relates to. Hence such discourse may indeed invite change in the product being discussed, and agreement be reached by some or all concerned (in that particular consensual event) that there has been a positive shift in quality.

Such discourse, such agreement or disagreement about quality, is itself a process of quality control, no less effective because it is collaborative, and no less effective because people disagree. As such it could provide another method of certification, as indeed it more or less does among the elite of any profession; a fact that for many would make a stronger case in this argument than any other. For example, the final educational judgment of this work is with two examiners, who may differ greatly in their opinions.

Standard products?

So what? If in measuring the standard we have denied what is essential in quality, does it matter? Lack of official recognition of originality, a little repression of creativity, is unfortunate but hardly crucial in the world order. Yet the other side of the coin may well be crucial in the order of the world. For what is involved here is not a single instance of non-recognition, but the very production over thousands of instances of the thinking person, of the learning person, of the person in work, of the person with authority; of, indeed, the moral, rational person.

For the standard is more than just one of many nudges and winks that lead the child to God. The standard, as applied continually through the strictures and structures of family and school and occupational work, at first externally and then through internal absorption and prescription, is the major mechanism, the quintessential carrot and stick, that moulds and shapes, that produces and creates that consciousness that defines the way each person sees the world, thinks about it, and acts within it. Not entirely, but largely so. And the individual produced through the notion of the standard, with its sharp cutting edge of adequacy, is a much more conforming, accepting, black and white, uni-dimensional person, and hence one far more socially controllable, than is one produced though the more spontaneous, multi-dimensional and unpredictable notion of quality.

Maybe we don't need to de-school. Maybe all we need to do is to acknowledge the arbitrariness and error that permeates standards and their measurement, extol the virtues of immeasurable quality, step lightly and quickly aside, watch the catagorisation structure crumble, have faith in chaos theory to articulate another structure, and hopefully nudge it in the direction of greater rationality and equity, truth and compassion. But that's another story.

Summary

The notion of the standard intervenes in the discourse about quality, and severely distorts it. The standard is a member of the class quality, is separated from it because of properties of measurement accuracy it is purported to have, yet is still confused with it. When the standard is seen, realistically, as unable to perform its function, we must return to quality as the notion with sufficient mythical, ideological, and intellectual status to replace it. This would predispose us to a rather different political structure, and to the recognition of a world in which simplistic notions of linear competition and dichotomous categorisations are replaced by more complex, ecological, and collaborative axioms.

Return to Table of Contents

Part 4: Error analysed

Chapter 13: Four faces of error
Chapter 14: What do tests measure?
Chapter 15: The psychometric fudge
Chapter 16: Validity and reliability

Chapter 13: The Four Faces of Error

Synopsis

The meaning of error in each frame of reference for interpreting assessments is now considered: In the Judges frame the phrase "error in the Judge's frame" is recognised as an oxymoron; in the General frame error is conventionally defined in statistical terms that ignore or underestimate some of the considerations, and the unattainable true score is seen to be a theoretical construct that need not relate to any external reality; errors are hidden in the Specific frame, and some of the Pretenders to this frame, namely mastery tests, criterion referenced tests, and competency standards, are briefly examined; finally in this chapter the meaning of error in the Responsive frame is considered. As this frame involves human interaction and discourse, error is what disrupts or disturbs movement towards clarification of meaning.

Assessment discourse is necessarily confused and confusing when the frame of reference within which the discourse is occurring is not specified, or when it involves definitions and methods where the actual frame being used is misrepresented.

The meaning of error in different frames

As soon as assessment data are committed to paper, their material permanency is dramatically increased. Likewise, the span of their associations is spread and emphasised. No longer just a description of a particular performance, the assessment becomes interpreted as a measure of knowledge and ability, an indicator of achievement on a course of study, and a predictor of future success or failure. Participation in an event has been transformed into an attribute of a person.

To estimate error is to imply what is without error; and what is without error is determined by what we define as true, by the assumptions of the frame of reference that forms our epistemological base.

There are four, at least, frames of reference for assessment. Four different sets of assumptions about the nature of the exercise. So within each of these frames the meaning of error, as defined by the assumptions of that frame, is different. Just as the meaning of error within each frame will be different again if judged by the assumptions of another frame. It is these differences that will be examined in this Chapter.

Error and the Judge

The Judge assumes omnipotence and infallibility within limits. The limits are defined by the particular performances with which the Judge is presented. These are the facts of the case. The task of the Judge is simple. He examines the performance of the accused, in whatever form it may be presented, he relates this performance to the standard, and then describes it accordingly.

He does this without error.

So problems that relate to error such as labelling, construction, stability, generality, prediction, categorisation, values and distortion of learning are, to the Judge, irrelevancies. For Judges are practical people, concerned with the realities, with what is, rather than what might be. And for them reality is the answers written on paper, is the art poster presented, is the motor repaired; in short, is the performance or artefact with which they are presented.

Questions of ability and stability, of looking to the past or to the future, are both irrelevant and unsettling. Irrelevant because they are outside the limits of their scrutiny. Unsettling because they trigger notions of a subject.

> *What sort of jargon is that?*
>
> *Is what?*
>
> *Trigger notions of a subject, for God's sake!*
>
> *You find that a bit obscure?*
>
> *I find that absolutely obscure.*
>
> *I was alluding to the difference between subject and object.*
>
> *I'm none the wiser.*
>
> *An examination paper is an object. A grade is an object. A standard is an object. The Judge relates these objects. And he claims to do it quite objectively. A computer, programmed correctly, would also do it objectively. Objectively in this context means that the process is purely rational, untainted by emotion or expectation of any kind.*

> *The Judge is firmly positivist in his stance; he rationally assesses what is out there in the real world to be described.*
>
> *Seems eminently reasonable.*
>
> *Indeed, if somewhat inhuman. An observer in another frame of reference might see the Judge as myopic and deluded. He might see the Judge immersed in a totally subjective world triggered by the statements, now confined to paper of the person being assessed. Further, he might see the comparison with the "standard" as an intuitive rather that rational process, affected by images, emotions and expectations stimulated by script, time and style of the answers as much as by content.*
>
> *That also seems eminently reasonable.*
>
> *Regardless, it is necessary for the Judge to deny such subjectivity in order to maintain the role of impartial expert, of perfectly calibrated measuring instrument. The Judge considers his work as objective, and so is unsettled by the notion of the subject, the four dimensional person who is assessing, and the four dimensional person who is being assessed*

Most teachers marking tests and assessing student work, and most public examiners, work within this frame. So most educational assessment is, by definition, error free.

Sometimes it is necessary, because of numbers of students, to have more than one Judge. There may be a number of Lesser Judges and a Chief Judge. In such situations it is accepted that ratings from lesser judges could contain some error, of the order of one or two marks in a hundred. To minimise this possibility, sample answers for questions might be prepared, with detailed marking shedules.

Sometimes a further check is made of papers just one or two marks below the cut-off points for failure. The Chief Judge will examine these to ensure that there has been no error, thereby restoring the myth of infallibility.

Reducing error in the Judge's frame of reference is not a problem. There is no error, except in the special cases of Lesser Judges and crucial decisions. In that case the error is the difference between the original assessment, and that of the Chief Judge.

Note that the Judge is infallible regardless of the form in which he presents the assessment. He may compare with the standard in any way he thinks desirable. The Judge is perfect in his rank orders, scores, grades, or other normative classifications. He is equally impeccable should he present his assessment in any other form, such as verbal description, moral tirade, or hologrammed logo.

The important point to understand is that the Judge is part of a social and political structure in which the inviolability and accuracy of the Judge's decisions are crucial elements. To suggest that the Judge may be in error threatens the stability of that structure and its accompanying mythology, so it is an act both treasonable and blasphemous: treasonable because it undermines the structure of society; blasphemous because it denigrates one of its icons.

In the hundreds of letters I have read in newspapers complaining about examinations, I have never seen one that suggested that the Judge, because he is a normal person, may make whopping big errors! So to the general public the Judge is not a normal person, and makes no errors.

Error and the General

Most of the book space and discourse time about this frame has been appropriated by those associated, corporately or academically, with the test construction industry; by those who produce and sell achievement and ability tests of many and varied kinds. Or by those who play in a scholarly way with mathematical models that might be used by those who construct such tests. (Nairn 1970). I shall deal with this world specifically in Chapter 15, the psychometric fudge.

Within this frame as constructed by psychometricians the error is the difference between the true score and the estimated score

However, the logic of the frame does not require such elegant and complex mathematical manipulation. The mathematical models have, overall, been counterproductive. Their theoretical elegance has hidden their inapplicability to most practical learning and teaching situations; the mystification of their statistical constructs has hidden from teachers, students and public alike the enormous extent of rank order inaccuracies and grade confusion, and the arbitrary nature of all cutoffs and standards.

One further point needs to be emphasised here. The General frame contains no notion of Standard. It is about creating stable rank orders of students. Anyone, anyone with sufficient authority that is, is at liberty to arbitrarily define a standard somewhere along that rank order. But a standard so defined is obviously a relative, not an absolute, division.

Error and the Specific

In this section we will look at error in the Specific frame in its purest form of specific behavioural objectives, as well as in its degraded states of mastery testing, criterion referenced testing, and competency standards.

In this frame there is only one correct description of performance, and that is the unambiguous learning outcome defined in advance. It is assumed that learning outcomes can be defined so clearly that there is no doubt about whether a student has, or has not, matched behaviour to objective. In such a situation there should be no problem with labelling error because there is no labelling. Each objective stands alone, pure and clear in its pristine self description; context, task and standard clearly enunciated. (Mager, 1962)

Construction errors are another matter. Whilst it is assumed in this frame that any outcome relevant to a particular course of study can be so specified, it is not claimed that all such relevant outcomes are in fact described. In some cases only those outcomes that all students are expected to attain are specified. Then we have a set of minimal learning outcomes. In asking "who makes this decision" we indicate a construction error. Why these particular objectives? And why these particular cut-offs for adequacy? It is apparent that behind the asserted certainty and objectivity of these objectives lies the usual minefield of idiosyncratic and arbitrary construction errors.

In other cases, a set of possible outcomes may be taken as indicators, and attainment of these is taken as evidence of achievement of related ones not directly assessed. And of course, no performance is ever a perfect indicator of a related performance, so hiding behind this wall of tightly specified objectives are all of the errors related to generality as well as to construction.

These construction errors, however, are all quite small compared to the massive one involved in the basic assumption of this frame: The assumption that any outcome pertaining to a course of study can be specified according to this frame; that all important outcomes can be specified in the form of a specific behavioural objective. In practice, it is just not so. This is what Messick (1989, p63) refers to as "construct underrepresentation".

This method of description is appropriate for situations where there are a finite number of tasks. Conceptually we are limited to tasks involving low level comprehension. As soon as we move into problem solving, analytic, application, or creative activities, there are an infinite number of possible task situations in which a student may be put in order to assess whether the student can demonstrate these more complex cognitive and practical operations. The tasks are limited only by the imagination of the test setters. And if we choose any one of these tasks, and describe them in such a way that they can be "taught" as a specific objective, then the task becomes one of low level comprehension. In other words, it must be a new task, a task previously unspecified, if these higher level performances are to be indicated (Bloom, 1956).

A student may attempt the task on a number of occasions if necessary, so usually irregularities in the performance of a particular student are not considered significant; unless, of course, a requirement of regularity over time is built into the objective. So errors in the temporal dimension are not applicable - unless, of course, we wish to infer that because a student has done the task, the student not only can do the task now, but on all occasions in the future. Such inferences are often made, of course. And they are utterly indefensible.

Prediction errors for an individual objective are enormous. But then, a specific objective does not claim that it would alone, or even in conjunction with other objectives, predict anything. On the other hand, as soon as it starts to describe itself with other adjectives, such as minimum, or essential, then it does open the way to predictive estimates of error.

Error and the Responsive

In the Responsive frame for any student there are many descriptions that are accurate and adequate to a particular purpose. Adequacy means that the description conveys sufficient information to carry the intent of the assessor and/or assessed into effect.

In this frame there is no competitive element, nor are the outcomes predefined in detail. Rather the assessor responds to the situation in terms of a particular purpose, which might be to describe how the student could improve the performance the next time (descriptive assessment). Or a responsive assessment might lead to a student's involvement in planning and assessing a course about maintaining a tractor (work required assessment). Or a responsive assessment might involve sharing a personal non-judgmental response to the student's work (detailed audience response).

While sometimes the criteria used for a responsive assessment might be preconceived, this is often not the case. The criteria emerge out of the totality of the situation, and so depend on the assessor's sensitivity, empathy and sense of quality. In addition, notions of adequacy are in general accepted for the subjective entities they indeed are, so become notions for considered opinion and discussion, rather than pretending to be absolute, accurately measurable qualities.

Responsive feedback then is part of a communication process which involves observation or other sensory input, interpretation, and response. It may in addition involve ongoing dialogue. Inaccuracy, in the sense of misinterpretations or misunderstandings may occur at any of these stages, as may obfuscations, denials, irrelevances, or contradictions. Empirically, this reduces to differences in interpretations, and there is no necessity in most cases to assume that there is some "true" interpretation or description. The aim is not to accept or reject the other's meaning, but to understand it.

In this frame, the person being assessed is also a potential observer and assessor, so self assessment can be an important part of the process. The communication process tends to be self-correcting, as the parties to the interaction both are concerned to clarify and understand what is being communicated. Accuracy then is concerned with the clarification of meaning, and error is reduced through openness of the communication channels.

Adequacy can only be determined by consequences. That is, to the extent to which effect conforms to intent. Again, error is reduced in as much as the assessed can feed back to the assessor the effect of the assessment, so that modification either of the description or the purpose can occur if necessary. This assumes that the assessed is aware of the purpose of the assessor's comments, and has reflected on their effects. So the continuity of open communication is as necessary as its initiation.

Keeping all communication channels open is of course more easily said than done, particularly in the social milieu that pervades most teaching-learning situations. For optimum reduction of error in this frame, both teacher and student would need to value openness over protection, autonomy over control, uniqueness over standardisation, complexity over simplicity, and tentativeness over certainty. In addition, each would need to be conscious of the potentially debilitating effects on open communication of the hierarchical structure in which their relationship is probably embedded.

More importantly, each would be wise to be aware of the potentially destabilising effects of their open communication on that structure, and of the social risk involved in such radical activity.

Summary

As the meaning of error changes with assessment mode, so do the methods designed to reduce such error. From a perspective of oversight of the whole assessment field, this is itself yet another source of confusion and invalidity, particularly as it is rare for any practical assessment event to remain consistently within one frame of reference.

Return to Table of Contents

Chapter 14: What do tests measure?

Preview

In this chapter I discuss in more detail the question of what it is that a test measures. In what sense can it be said to measure knowledge or ability? To what extent does it perform a ritual task and measure nothing? Or is it the wrong question? Should we rather ask, what do tests produce?

Tests and scales

A measure, or scale, assumes of course that equal intervals anywhere on the measure are in some sense of equal value. That the difference between sixty and seventy percent is in some way equivalent to the difference between twenty and thirty percent. So if a test is a measure then it must be a measure of something, and we would expect equal differences to represent equal differences in that something.

We know that a ruler measures length and the unit is a metre. We know that a clock measures time and the unit is a second. We know that a balance measures mass and the unit is a kilogram. And relative humidity measures what fraction of the water vapour the air could contain at a given temperature that in fact it does contain. So this is a pure number. Nevertheless, it is a ratio of two quantities that do have units.

So what does a test measure? And what is the unit of measurement? Let's look at the unit issue first.

It is clear that there are no units. The measure is a pure number. Unlike relative humidity, however, it is not a ratio of two measures of absolute humidity which do include units. Again, this supports the idea that the numbers are not measures, but ordinal numbers - numbers that represent an ordering of some kind. Numbers that describe a position in a series. Numbers in this case that assert that some performances, or people, have more of "something" than do others.

At this point it is worth mentioning that the whole paraphernalia of normalising scores and otherwise fiddling with them has two purposes: One is to try to magick a linear scale out of an ordinal one by making various sorts of assumptions about the distribution of the "something" that is being "measured"; the second is to produce "measures" that are mathematically pliable, that are accessible to the manipulations and pleasures of mathematicians; that will, in short, turn a horse race into a profession (See chapter 11).

Cultural differences

Back to the problem of the "something" that is measured by the test. For the most part, Europeans and their colonial converts on the one hand, and the United States and their spheres of influence on the other, have different approaches.

To the Europeans it has never been a problem. Inured by tradition to a religious belief in the Judge they have generally accepted the proposition that the test or examination measures whatever the Judge says it measures. The acceptance of this "fact" denies the existence of a problem. The Judge says that tests measure student achievement. Pressed further, he or she might say that student achievement is a measure of what has been learnt on the course of study being tested. The test is simply that part of the course where learning is demonstrated. And the Judge, who holds the mystical secret and truth of standards, is able to convert this demonstration into a mark which is the true measure of what is achieved.

As I wrote that last paragraph I was aware of how "right" it sounded. Like all religions, there is a plausibility in its logical circularity that is terribly enticing, a simplicity in it's self-evident truth that gives a deep sense of security. Articles of faith are characteristically immune to both the challenges of logic, and the intrusion of empirical data. To paraphrase Horkheimer and Adorno (1972), faith needs knowledge to sustain it, and thus pollutes knowledge in the act of attaining it (p20).

The Americans, whose religious tradition is democratic and competitive rather than monarchic, have little faith in particular Judges or, for that matter, Presidents. Which is not to say that they do not revere even more in compensatory manner the institutions of power in which these fallible humans are niched. Regardless, their tests must be free of the Judge's subjective idiosyncrasies, and pay due homage to the competitive individualism that is central to the American dream.

The problem of subjectivity was (mythologically) solved through the medium of the "objective" test:

> The major premise of the American system of social morality is that every individual should have an equal opportunity to compete for the prizes offered. . . that every contest be objectively judged, as impersonally as possible, with no favouritism, nepotism, or any other kind of ism. To make this objectivity evident, access to preferred categories should, wherever possible, be granted on the basis of scaled scores that a machine can handle. (Friedenberg,1969, p28).

This has the added advantage, of course, of being "economically efficient," another central tenet of the American dream.

So the "something" that the test measures is measured economically and objectively, but we are still left with the sticky problem of what this "something" is. For when the Judge goes away, this problem raises its (previously covert) head.

Over the years, American test gurus have come up with a plethora of things that they claim to be measuring; intelligence, specific ability, attainment, achievement, competence, factors of the mind, specific outcomes, curriculum objectives, minimal competencies, true scores, universe scores, latent traits. An interesting oscillation between physics and metaphysics, between outside behaviours and inside mind-potential, between performance and hypothetical mental structures. Be assured however that efficiency has been conserved. In many cases the same test item can be used to measure all of these "things." (Nairn 1980; Taylor, 1994; Sternberg, 1990)

The simplest conclusion is that multiple choice tests measure exactly what the people who construct them claim that they measure; the definition of the abstraction they claim to measure is simply the score on the test. Which puts the Americans in a similar position to the Europeans, with the substitution of test agencies for individual Judges, of an elitist junta for the monarchy.

One corollary of this conclusion is that the tests really do measure something but no one is sure what it is. In the light of all of the evidence this seems unlikely to me. Contradictions are predictable from the logical type confusions that are inherent in the whole test process.

A more plausible corollary is that the tests do not measure anything in particular, nor do they place people in any particular order of anything, except the order that participating in testing events of any sort tends to generate. But they do place them in an order, along a single line of "merit," and that is all they are required to do.

One more point is very significant. "Ability" or "achievement" tests like the Scholastic Aptitude Test do place groups of students (not to be confused with individual students) in an order very closely related to parental income and social class. In this sense they contribute significantly to the stability of an unequable social structure whilst at the same time producing an ideological smoke screen by asserting that they are ordering on the basis of individual ability. And the victim pays for the test. Fantastic! (Nairn,1980; Friedenberg, 1969, p29).

Social skills

In 1976 I was about to begin a five year research project looking at social development in school classrooms. At the time there was much educational discourse about teaching social skills, which many thought were in short supply in young people. "Improving social skills" was an objective in courses from grade one Mathematics to grade seven English to grade twelve Economics. As part of the preliminary work I visited schools in Australia, Canada and the United States, and talked to many teachers about the social development of their students.

These teachers were all interested in the social skills of their students. They taught young people from the age of five in infant schools to the age of seventy five in Ph.D programs. Yet in describing their students to me there was enormous similarity in their descriptions. It went something like this: "When they first come to me they are pretty bad. Inarticulate really. Stumble over words, tend to answer just yes or no. Can't put two coherent sentences together. Can't listen properly. Can't concentrate. Just don't seem to be able to relate to other people. Bad with their peers, and worse with me. Then as the year goes on and they get more practice in speaking up and their confidence grows they improve tremendously. By the end of the year I've generally been able to produce a class with quite mature social skills."

What particularly struck me about these conversations was that they appeared to be the same regardless of the age of the students. So how could the social skills of five year olds be the same as those of twenty five year olds?

Then I thought about my own experience over the previous two years as a "leader" of communication workshops; thirty teachers doing residential five day courses to increase their communication skills. Weren't they exactly the same? At the beginning of the week hesitant, not really listening to other people, insensitive to feelings. Then by the end of the week attentive and empathic, talking poetry rather than cliches.

Had we been asking the wrong question? Did this change have anything to do with learning new skills? Or had we, over the five days, changed the social environment so that it was now appropriate to engage in a different sort of dialogue? Had the group experiences produced enough trust and cohesiveness to allow for some flow in human relationships, to overcome the stultifying role restrictions and mistrust that characterise much of our normal discourse? Were these observed changes simply indications of emotional openness, with concomitant increase in divergent thinking and spontaneity?

The implications for our research were clear. The question we should address was not "How do we teach better social skills?," but rather "How do we develop the classroom group so that mature social relations and discourse are appropriate?"

How can a social skill belong to one person? At least two people are always involved, and what is appropriate interaction, whether verbal or non-verbal, must always be a function of the relationship between them, of the context of the communication. What appears to be a quality of the person, a skill, turns out to be a production of a particular environment, a particular aspect of a human interaction, a discourse appropriate to a social relation.

As with the quality of the bridge, so with the quality of social behaviour: Even if it can be labelled, the label can't be pinned on any particular object.

Knowledge

> *You rigged it.*
>
> *What do you mean, I rigged it?*
>
> *You wanted to prove your point about not pinning a label to a person. Then you chose social skills to talk about. And OK, you've got a case there. But what about intelligence? What about intellectual skills? What about cognitive achievement? What about mental ability? That's where the action is.*
>
> *Certainly that's where the money is. Skills are what employers seem to want, and increasingly what education seems to be about. And as you suggest, cognitive skills, facts and knowledge and understanding, are at the high status end of the skills spectrum. But why are they so different to social skills?*
>
> *Because they surely do belong to a single person. You don't have knowledge in relation to someone. Analytic ability is not a relationship with another person. Reasoning skills are*

surely inside the person and not in some mystical relationship that characterises an event.

So let's look them in turn in more detail. Let's take knowledge first. If it's knowledge we're talking about, then it's got to be knowledge about something. So choose something.

Computers.

So how would you know that you had some knowledge about computers?

I've used them at work for various things; cataloguing, letters, drafting. So I know what programs to use for particular purposes, and I know how to use them.

In other words, you would reflect on particular interactions that you have had with computers, and on the results and feelings associated with those interactions?

I suppose so.

And you would interpret that recall of those experiences as knowledge?

Well, if I hadn't had the knowledge I couldn't have done the work.

But you just told me that you only knew that you had the knowledge because you had done the work.

Yeah, well that's now. But what about the first time?

What about the first time?

The first time I must have had the knowledge first or I couldn't have used the computer properly.

Tell me about the first time. Did you use the computer properly?

Well, you know. I had to mess around and experiment a bit before I got it right.

So the first time you had some knowledge, but not enough to do it properly?

Yeah.

And how did you know that you had enough knowledge to *even make a beginning?*

Well, that needed a bit of confidence, and a bit of taking a risk.

So it required a certain emotional state as well as a little preliminary knowledge?

Yeah, that's right.

And how did you know, or suspect, that you had that preliminary knowledge?

Well, I'd done some other work with computers. And of course there was the instruction manual.

In other words, you recalled other experiences with other computers. And you followed the instructions in the manual.

So is the knowledge in the instructions?

The instructions are meaningless without an event involving an interpreter and a computer.

Ok. If I had to follow the instructions then I didn't have enough knowledge. Reading the instructions became part of the event and enabled me to proceed. Now they are part of my experience that I can recall for future events.

So knowledge, once again, becomes, or at least involves, the process of recalling prior interactions.

So you reckon my "knowledge" of computers consists of reflections about real past events, or following instructions to produce an event which I can recall, in which I interact with a particular computer in particular ways. Knowledge appears in this case to be the construction, or the reconstruction, of an interactional event, a relational experience. Knowledge also implies that the emotional tone of that event is positive.

Exactly. Knowledge isn't something that you have. It's something that you do. It's something that is reconstructed in the present from memory traces of things that you've done before. You can carry out those reconstructions visually or in language in your own head, or in action with whatever objects are involved. And so knowledge of a particular field is continually created and recreated in the processes of selecting and applying memories of experience in that field.

Stories

Let's make a slight diversion here to consider how this process of learning occurs developmentally in young children.

> All children, roughly between two and a half and four years of age, start to comprehend and make up narratives about their own lives. Also, adults of all cultures express their history, beliefs, values, and practices in the form of stories as psychological narratives. These stories are among a culture's most potent forms of self-expression and among its most effective forces for perpetuating itself (Stern, 1991, p133).

By creating a story, we create a reality. And we have as many realities as we create separate stories about ourselves in the world. It is in the creations of such stories that we define ourselves to ourselves. Out of our past we select and choose the experiences, with appropriate perceptions, that sketch the outline, and then fill the substance, of our stories. The firmer the story line becomes, the more selective our experience, and the more distorted our interpretations are likely to be, to maintain the story line. All this is fine, so long as we keep reminding ourselves that we are much more than our stories, that our experience is much richer than our perception and interpretation of it, and that the world is much more than our experience of it.

Yet there is another trap more subtle still. For not only do we get caught up in our own stories, we also get caught up in the stories of other people, particularly those we admire, or love, or are controlled by. For we do not live alone. We are social animals, and our life stories require other people to bring them into being.

Thus our stories about ourselves in the world are constructed out of our experience in the world. And this experience may come to us by direct involvement in the world, or involvement through the incorporated stories told us by others. And once these stories become accepted by us, they become part of our reality, part of our way of living in the world. Then we tend to construct our experience out of our stories. This is not a cause-effect relation, but an ecology of effects; our consciousness of the world, our way of being, involves an intimate interconnection of our experience, and the stories we use to make sense of that experience.

Our knowledge of ourself is just that interconnection.

Knowledge of a field

In just the same way do we construct knowledge in a particular field of study. We create events around the object of study, observe what happens, and then make up a story about what is happening. Or more likely accept someone else's story about what is happening. For any field of study is just such a consensus story, comprising what Foucault calls a "regime of truth." Then we use the story to help us make sense of other events involving the object, or other objects in that field.

This is equally true whether the field of attention is immense, as in mysticism or physics or history or engineering, or is small, as in building a table or washing dishes or driving a car.

So our knowledge of the field consists of descriptions of events involving a selected set of data constructed out of the relation between story and experience, between hypothesis and interpretation (possibly involving measurement), between conception and perception. As Wolf (1991) expresses it, "sophisticated thought follows a 'zig-zag' course between craft and vision"(p41).

But again, let us be clear on this fundamental point. The data, the knowledge, does not belong to the object of study. It is not a property of the object. Nor is it the name or a measure of a property of the object. It is rather information about the relationship of the object to its environment during a particular event, a particular interaction, suggested by the story in which it has a part to play.

Messick (1989a) comes close to this but does not follow it up. In claiming that tests "do not have reliabilities and validities, only test responses do," he goes on to say "that test responses are a function not only of items, tasks, or stimulus conditions but of the persons responding and the context of measurement" (p14). In my terminology, they are functions of events.

We could generalise. All knowledge is knowledge of the relations that identify events. And as we are observers at some point in the interaction, either at the level of direct observation, or at the level of constructing and interpreting the story that is the basis for the data collection, then we ourselves are involved in the interaction, and are thus part of the knowledge. And for the very reason that we are part of the knowledge, we are not that knowledge, and the knowledge is not part of us.

Human ability

In the light of the above, how are we to make sense of the notion of human ability, of capacity, of intelligence, of cognitive achievement, of some factor of the mind, of a latent trait?

These are normally considered properties of the person, attributes of an isolated mind, functions of an individual human consciousness. Yet our analysis of how we collect information about the other, or even how we obtain knowledge about our self, denies the possibility of such separation, and acknowledges the possibility only of information about relations.

I described knowledge of the field as a selected set of data constructed out of the relation between story and experience. Such selection is always in a context of some action, even if the most recent action is talking to oneself. Ability is a redundancy concept that acknowledges the action and then claims responsibility for it. It is an example of the common epistemological error of attributing a cause to the relational balance of an ecological system.

Semantically, this is achieved through the simple trick of nominalisation; of changing a verb into a noun, and thus of converting a process into an object. It is very simple: I do something, I am part of an event. Therefore, the causal logic goes, I am able to do the things I do (before I do them), otherwise I wouldn't have been able to do them. Therefore I must have (here comes the nominalisation) an ability, some property located somewhere within me, that allows me to do this thing that I do.

This is an example of the dormative principle. Keeney (1983), explains how it works:

> To invent a dormative principle, begin with simple descriptions of the phenomena to be explained. For example, a person may be described as unhappy and unwilling to work or eat. These descriptions can be classified as a category of symptomatic action such as 'depression'. The claim to then 'explain' these particular descriptions as the result of 'depression' is to invoke the dormative principle. What one does, in that case, is to say that an item of simple action is caused be a class of action. This recycling of a term does not constitute formal explanation.(p33)

The fact is the action: I run, or I try to run and cannot. What happens when the "ability" construct is introduced into the story? Now the reason I can run is that I have the ability to run. My running has a cause. I have some permanent property, some palpable attribute, that accounts for my running. My running is no longer a dynamic process of relationships between muscular and visual coordination, of memory and environmental feedback. My running is no longer a variable dynamic. It can be described as a causal relation independent of time.

My running is now explained by a little permanent stable packaged bundle of something inside me called "ability to run." It is a fixed static. As such it is a glue that helps fix me in time and space. It enables me to be compared, labelled and classified in terms of this property. It becomes part of my individuality.

What difference does it make? It makes world of difference, and a difference in the world. If the limits to my occupational choice and political power are largely determined by my cultural experience, by my practise in the field in which my interest lies, then most people might legitimately claim the right to such experience.

On the other hand, if my ability severely limits my possibilities in that field, then I have no legitimate claim to further practise. My exclusion is legitimised. I cannot become a doctor or engineer or lawyer not because of lack of opportunity or experience, but because of lack of ability.

Foucault (1992), in two condensed epigrammatic passages, sums up the essence of this argument:

> The individual is no doubt the fictitious atom of an 'ideological' representation of society; but he is also a reality fabricated by this specific technology of power that I have called 'discipline'. . . . power produces; it produces reality; it produces domains of objects and rituals of truth. The individual and the knowledge that may be gained of him belongs to this production (p194).
>
> . . . the disciplines characterize, classify, specialize; they distribute along a scale, around a norm, hierarchize individuals in relation to one another and, if necessary, disqualify and invalidate (p223).

It is not by accident that whenever universal education claims to equalise opportunity to cultural immersion and hence occupational choice, at the same time examinations and psychological labelling provides upper limits previously applied through the mechanisms of class and caste. The basis of the highest morality of any society has always been the maintenance of stability.

Conclusion

So what does a test measure in our world? It measures what the person with the power to pay for the test says it measures. And the person who sets the test will name the test what the person who pays for the test wants the test to be named.

The person who does the test has already accepted the name of the test and the measure that the test makes by the very act of doing the test;

when you enter the raffle you agree to abide by the conditions of the raffle.

So the mark becomes part of the story about yourself and with sufficient repetitions becomes true: true because those who know, those in authority, say it is true; true because the society in which you live legitimates this authority; true because your cultural habitus makes it difficult for you to perceive, conceive and integrate those aspects of your experience that contradict the story; true because in acting out your story, which now includes the mark and its meaning, the social truth that created it is confirmed; true because if your mark is high you are consistently rewarded, so that your voice becomes a voice of authority in the power-knowledge discourses that reproduce the structure that helped to produce you; true because if your mark is low your voice becomes muted and confirms your lower position in the social hierarchy; true finally because that success or failure confirms that mark that implicitly predicted the now self evident consequences. And so the circle is complete.

Return to Table of Contents

Chapter 15: The psychometric fudge

Synopsis

The first part of the chapter details some of the ways in which psychometricians fudge; by reducing criteria to those that can be tested; by prejudging validity by prior labelling; by appropriating definitions to statistical models; and by hiding error in individual marks and grades by displaced statistical data, and implying that estimates are true scores.

In the second part of the chapter a number of specific examples of fudging are detailed; in particular, the item response theory fudge, selection and prediction fudges and the great Queensland reliability fudge.

Constraining the definition

Reliability and validity are two concepts dear to the heart of test constructors and others involved in the field of psychological and educational measurement. I'll begin my analysis of the fudge that characterises the field by looking at reliability, or the lesser fudge.

Reliability in classical test theory is (indirectly) an estimate of the error you'd expect if the student did a hypothetical parallel test. And in generalizability theory it's an estimate of the difference between the "universe" score and the score on any particular test. In both cases it's about the reliability of the test, or more accurately of the test-testee interaction, and not of the assessment; of the extent to which two tests give the same score, not the extent to which this particular description of student performance, based on a test, confirms or contradicts other such descriptions, which may or may not include a test (Behar, 1983, p19).

Note the way the mathematical model simplifies and constrains the world. It would be easy to believe the reliability of the test was about the extent to which the test describes course outcomes or student performance or work successfully completed. It isn't. It confines itself to the closed world of the test. It's about its ability to reproduce itself.

Mathematical models and true scores

The concept of the true score or universe score is central to the derivation of the theory. That is, it is a theoretical assumption. That does not mean that it necessarily has any place in the interpretation of the theory, that it corresponds to some measurable property of real people. And even if it does, the theory indicates that we can never know the true or universe score, only an estimate of it. And that estimate is always associated with error.

So in practice, in the world out there, there is no true score that can be attached to a person or an event. There is no thin line beside which a number is placed. Even before the empirical evidence starts to come in, there is only a wide fuzzy band, and all we can say mathematically is that the true score is probably in there somewhere. And if it is only probably in there somewhere, then for all practical purposes, for an individual person it isn't in there at all. In practice there is no true score. There is no stable rank order. And if in practise there is no stable rank order, then there can be no stable practical standard.

The history of achievement testing represents an enormous confusion of theory with practice. A model is not true or false. It is useful in as much as its predictions accord with empirical data at some points. It is not necessary that the assumptions of the theory correspond to actual situations in the world in which its predictions are applied. The assumptions of quantum mechanics from which the theory derives cannot be validated empirically. That is why they are assumptions. The metaphor in which the assumptions may be enclosed is useful in as much as deductions from the theory are experimentally verifiable. But such assumptions are not considered "true." Nor are they considered as having some "real" existence out there in the "atom."

Psychometricians on the other hand assert that their assumptions about a true score or universe score imply that such a score refers to some attribute, some measurable property, of a person. The person can be then classified, because the number is a measure of something called achievement, or ability, or whatever. In Criterion-referenced tests it is achievement in a specified "domain" of knowledge, and is called a "trait."

Regardless, this achievement is assumed to be some psycho-cognitive state which can be accurately described by finding a corresponding point along a one dimensional scale.

Why are these very intelligent people wanting to insist that their theoretical assumptions are consistent with empirical reality, when theories in general require no such correspondence? And when the fundamental assumption, the primary axiom of this particular theory, is that such correspondence can never be achieved? Why this enormous urge to represent uni dimensionally a variety of human performances which are obviously multi dimensional? Why this obsession with numbers, this illusion of numerical accuracy, this delusion of descriptive adequacy?

At this time, let us merely note that all of these activities are related to a psychological ideological assumption about human ability, or skill, or achievement. Some particular quantifiable quality of people that belongs specifically to them, and is thus independent of gender, race and class; that is unsullied by environmental factors; that is a permanent fixture of the person independent of the conditions of its production. That is, indeed, the clinging legacy of the nineteenth century belief that "intelligence was a unitary and immutable trait. It had no kinds or varieties, only ranks."(Wolf, 1991, p36).

As well these assessment activities are related to an ideological social assumption that this quality may be quantified and be represented along a uni dimensional line of almost infinite length, along which each person may now be accurately placed and categorised, their place permanently fixed, and their relative position in the order of things firmly established. And this conception of "ranking, fixedness, and predictability provided the "scientific" basis for two enduring institutional responses to the diversity of styles, cultures and academic backgrounds of students: universal testing and the systems of tracking students." (Wolf, 1991, p38).

And, further to Chapter 4 , note that

> This portable cumulative record of individual worth and achievement is central to bureaucracy and psychology alike. . . the inscriptions in individuality . . . make the individual knowable, calculable, and administrable, to the extent that he or she may be differentiated from others and evaluated in relation to them. . . individuality has been made amenable to scientific judgement. . . With psychometrics the previously ungraspable domain of mental capacities has opened up for government. What can now be judged is not what one *does* but what one *is* (Rose, 1990, p140).

The General frame and the true score

The logic of the General frame does not require any notion of a true score. The true score is a statistical artefact, a mathematical artifice, devised to defend a quite fantastic and monstrous proposition about ordering and classifying with great accuracy large numbers of people. Here is that monstrous proposition spelt out in more detail.

The political proposition that is being rationalised, justified, mystified, constructed and implemented in the notion of a true score is this: that it is possible in any area of human achievement to produce an accurate order of merit of "ability" in that area, and to attach to each person a number, a score, that fixes them firmly in position within that hierarchical order.

What do we actually know empirically? That under certain conditions it is possible to increase the stability of the rank order of merit of people on "test" results, in "test" situations. And that the more we can eliminate personal idiosyncrancies of setters and markers by averaging, and the shorter the time span of repeating the testing, the more the rank order is generalizable to other setters and markers of similar tests constructed by similar people.

We do not know empirically whether there is an asymptotic limit to this stabilisation; theoretically, and practically, there is always an error of measurement. We do know that this fits empirical data quite closely in regard to sampling assessors for marking. That is, when students do very similar tasks and the idiosyncrasies of assessors are "averaged" out.

We do not know empirically whether a similar stabilisation occurs when results are averaged over different occasions. There is no a priori reason to believe that they should be, especially for achievement tests with a high memory component. Indeed, there is every reason to believe that the actual performances of particular students would vary considerably, and differentially, when assessed over time, given that their forgetting curves are non linear and of different shapes. Thus sampling across these dimensions could produce an increase in error in the General frame, not a decrease. It would be very dangerous to collect such information, however, for it would contradict the assumption of stability that the notion of skill or ability implies.

Empirically the true score is not known, and can never be known. Empirically estimates of the true score can be obtained, and these are always different, because all of the measurements we make contain an error. In practice then, error is indicated by the difference between estimates, not between estimates and some hypothetical "true score." That is why the notion of true score is not necessary for simple and specific and individualised estimates of error, though theoreticians and ideologues may well require the idea for their own particular purposes.

The notion of the true score, then, despite its enormous ideological importance, is practically unattainable, irrelevant, and misleading. It is a theoretical input to the mathematical theory of testing, not a practical output. The statement that there is a true score is a statement about a theoretical statistical assumption, not about an attainable empirical reality. Further, such assumptions of mathematical models need have no direct links to any properties or aspects or qualities of phenomena "out there" in the real world.

> Note that we do not define true score as the limit of some (operationally impossible) process. The true score is a mathematical abstraction. A statistician doing an analysis of variance does not try to define the model parameters as if they existed in the real world. A statistical model is chosen, expressed in mathematical terms undefined in the real world. The question of whether the real world corresponds to the model is a separate question to be answered as best we can (Lord,1980, p6).

Lord then agrees with me, at least on page 6. More of Lord later. For now, having seen how the fudge about the true score works, we'll examine some of the others. One really big one relates to test items.

Models and items

There is no doubt that one way to get information about achievement (what a person has done), or skill (what a person can do), or ability (what a person could do given the opportunity), is to get them to answer some questions about what it is they are supposed to have achieved or have the ability in. And one rather contrived way of doing this is to use pencil and paper tests. Further, a particular method of this technique is to use test items of a multiple choice or short answer form.

It requires an enormous suspension of rational thinking to believe that the best way to describe the complexity of any human achievement, any person's skill in a complex field of human endeavour, is with a number that is determined by the number of test items they got correct. Yet so conditioned are we that it takes a few moments of strict logical reflection to appreciate the absurdity of this.

Test items not only determine the form and media of testing as paper and pencil tests, but also specify the type of question as short answer or multiple choice. In other words, talk of test items tends to narrow dramatically the sort of performance situation in which the person being assessed is to be put, and also severely limits the sort of description that might be given.

Why is this important? Because psychometricians have defined reliability and generalizability in terms of test variance, which is in turn determined by the characteristics of test items. Likewise, estimates of construct validity, on the rare occasions they are estimated empirically, are determined by statistical manipulations of item characteristics.

By appropriating terms like reliability and generalizability and validity, and defining them in terms of the mathematical properties of particular tests, professional test agencies and examining institutions perpetrate another grand fudge. These concepts become narrowly construed as properties of tests, or relations between numbers, rather than as useful criteria on the basis of which concerned people may judge the whole assessment exercise.

Item response theory and the absolute scale

Item response theory allows us to construct a scale in the same way that classical test theory and generalizability theory enables us to construct a true or universe score.

The magic is in the word "construct." It is theoretically constructible, not empirically constructible. In fact, the theory determines that the scale is absolute but improbable; the actual scale produced measures the probability (or if you prefer, the improbability) that any person to whom the scale is applied actually has that reading on the (theoretically) invariant scale that the theory constructs.

Just as objective tests are highly subjective instruments in which the marking can be done objectively, but it is implied that the assessment is objective; and just as the true score can never be measured but it is implied that the estimated score is that score; so the invariant scale of the criterion referenced test can never be physically produced, but it is implied that the test produced contains that scale, rather than its very error-prone physical manifestation.

Criterion referenced tests

Criterion referencing, as applied by professional test agencies, is not directly referring to course objectives or to student learning. Criterion referencing refers directly to test items. A criterion referenced test is one that is proscribed by tight delineations of the structure of particular tasks to be included in the test.

Advocates of criterion referenced tests often claim that the performance on such a test is judged in relation to an absolute rather than a relative standard. That is, that scores on criterion referenced tests are measures of achievement in a particular domain and do not depend on relative merit, but are informative in their own right.

This claim is another psychometric fudge. Criterion referenced scores are in no way absolute scores. They are norm-referenced. The norm-referencing is done prior to the test construction process at the item level, and not at the total test level during a specific application of the test. (Behar 1983, Glass 1978)

Criterion referenced tests contain all of the errors of Mastery tests plus one additional labelling error of great ideological significance. A sub-group of tests in this area, called sometimes Domain referenced tests, have developed a whole theory based on test item characteristics, which is very efficient. Efficient in the sense that students can be tested with less items than in the random sampling model for the same error (an error which, as usual, is never attached to individual scores). This is achieved by using known levels of difficulty of the items (based on random or other specified population estimates), in computing the student's score.

Nothing wrong with this of course. Except the labelling claim that these scores are absolute measures of a "latent trait." What is a latent trait? It is some "hidden characteristic" which some students have more of than others, and which is measured by the test. And those who have more of it are more likely to be able to answer correctly the more difficult items.

As all of the items in a Domain referenced test relate to some particular area of learning, such as reading comprehension, or computer skills, or simple calculus, or newspaper editing, or social skill, or whatever, then it doesn't really matter what "latent trait" means. The assertion that "it" can be measured absolutely is what constitutes its ideological power. Here is the ultimate rationalisation for intellectual and social stratification. Here is the number that describes each person's place on the continuum of ability or skill or whatever for any label that testing agencies wish to attach to the domain of items.

On the surface, of course, it is the specific label that assumes social importance. The claim being made, or at least strongly implied, is that such a test is an absolute measure of reading comprehension, or computer skill etc. But in focussing on the label, we are likely to miss the frightening significance and ideological sleight of hand that produced the "latent trait" as some substantive property or quality permanently attached to the person tested, somehow magically unrelated to the highly subjective, contrived, interrelational world where a student sits at a desk, reads some questions, and places ticks in computer marked boxes.

Such tests construct current fashionable truths. They are being presented as the latest panacea for testing human ability, or "skills" or "competencies" as they are now called; they are being presented as the theoretical support for an invasion of competency based assessments in all areas of human measurement (in schools, businesses, bureaucracies, or where-ever else hierarchies operate). So we should be clear about three things:

The first is that constructing a domain referenced test and naming it produces no evidence that the tests measures any sort of trait or ability that can be attached to an individual person (Lord, 1980).

The second is that they are not absolute, or error free measures; the scores are related to relative merit, and there is no "standard" performance or score that relates to any minimum or other grade of "competency" that can be theoretically attributed to any score (Glass, 1978).

Which takes us to the third point, which is a logical conclusion from the previous two. Domain referenced tests can make little contribution to a field of "competency" assessment which purports to describe (or more significantly measure) some "standards" of competency in various "skill" areas of human performance.

Limiting constructs, limiting error

Let's examine briefly how some of the more general criteria of assessment; labelling, construction, stability, generality, prediction, tend to be limited to what can be controlled by test makers.

Labelling is achieved by the simple act of giving a name to the true, or universe, or latent trait, score. Which means, in practice, to the estimated score. The errors implicit in the communication of what that label means, between those who define the course, those who teach it, those who produce the test, those who do it, and those who consume its product, are thus not considered. All of these people will give their various meanings to the label, and make their judgments accordingly. We may be certain that these meanings vary considerably. How much they vary will probably never be known, because it is not in the interests of any institution to uncover yet another source of error. Labelling errors are not currently considered in any estimate of test error. I believe they are immense.

If communication is its effect, then such confusions are, to the student, irrelevant. To the student the meaning of the label is the grade or the mark attached to it. Within the structure that contains the assessment system, the meaning of the label, as distinct from the meaning of the mark, amounts to little more than ideological gossip.

At least some students recognise the meaninglessness of the label. I remember vividly a television program which followed the fortunes of four students through the final months of their preparation for the University Selection Examination in New South Wales. One student in particular, a science student, a paragon, studied hard and reaped the ultimate reward. Straight A's.

Just after he received his results he was interviewed for the last time. He was obviously pleased with his success.

> "I suppose," the interviewer said, "this will be very useful to you in the future."
>
> "The marks?"
>
> "The understanding. The knowledge."
>
> "Oh that. No, I don't expect that to be of any use to me at all. I'm going to be a lawyer."

Likewise, construction errors are not estimated; they do enter the theoretical psychometric definitions of validity, but are in practice neither measured nor estimated. The major task of matching objectives to assessment to performance is assumed entirely by the test maker, and most of the errors within this activity are also disregarded, as easily as the errors caused by differing forms of assessment, and use of media other than reading/writing, which don't fit the format of test items on paper, are disregarded. It is assumed that the test is indeed contracted, and the performance required by the student indeed matches, the objectives of the course, or the criterion definitions of the test. Sampling processes that are used, even in professional testing agencies, are at the best primitive, and at the worst nonexistent. This part of test construction is nicely described as an "art" rather than a science (Nairn ,1980).

One thing is certain though; no course has stated as its major, or even minor objective, the ability to answer a pencil and paper test in a given time under stress conditions. And why not? Surely this is the essential behavioural objective.

Stability becomes narrowed to test reliability, more accurately called internal consistency, an internal test measure that cannot take account of variation over time and place and assessors. Theoretically test-retest reliability is one form of reliability, but in practice such estimates are rarely obtained.

Generality becomes narrowly construed as related to the extent to which the test samples the universe of possible test items, or how well the item specifications cover the domain. Generality becomes a function of test items and is called generalizability. Generalizability ignores previous performance in different contexts, forms and media. It ignores all performance other than the purely cognitive response to simulated experience of a multiple choice or written form. It thus ignores all cooperative and all production modes of expression. It reduces human response to the act of recognising a "best" answer, to conforming adequately to some authority's view of importance, relevance and reality, or to answering someone else's question in a particular way.

And prediction becomes tied to numbers and test scores. In this psychometric world we are no longer concerned with the extent to which actual people are helped to function in differential social situations of great complexity. Prediction does not attempt to describe the relationship between a particular set of learning experiences for some person, and how helpful that is in some future situation for that person. Rather it ranks a group of people on their "success" in the "learning" situation, then ranks them again in some criterion situation. The correlations between the two rank orders represents the predictive value of the test. Not of the course, of the test. And not of its relevance to the quality of their performance, but to its correlation with some person's or group's ranking of their relative performance. And note that even if this correlation is high, which is unusual unless a similar test has been used to measure the criterion, this tells us nothing about whether the relation is in any way causal.

How the fudge works

The psychometric fudge occurs through the following processes:

Firstly, the criteria by which assessment is determined are chosen so that they are easily adaptable to the construction of tests and to the statistical manipulation of test data. Criterion-referenced tests are just that: Only those criteria that are appropriate for referencing test items are chosen.

Secondly, the validity of the test is prejudged by labelling it to describe what it is supposed to measure. Such is the power of labelling that this exercise in wishful thinking, this untenable assertion, is interpreted by most people, including the test constructors who become entranced with their own propaganda, as being an accurate description. At a deeper level still the mathematical theory itself contains such terms as true score, ability, and trait before any empirical information at all is available; that is, before any connection (let alone correspondence) with the world outside mathematics is established.

Thirdly, definitions are appropriated and defined to fit specific statistical models; in particular, by narrowing the universe of possible test situations to a universe of possible test items (random sampling model), or by narrowing the universe of possible test items further to the universe of suitable test items (domain referenced testing). In both cases the performance of students outside of such test situations is disregarded, or downgraded, and the right to appropriate the personalising labels (ability, trait, true score) is assumed.

Fourthly, the data is presented in a way that is misleading at best and deceitful at worst, by hiding error of individual marks and grades with obscure and displaced statistical data, thus implying, to all but the statistically sophisticated, that estimates are "true" scores. Further, the implication is made that such tests are accurate as predictors, claims that in most cases cannot be substantiated (Reilly, 1982). Finally, estimates of confusions and errors related to construct validity are ignored, usually theoretically, and almost always practically.

We could look at these fudges as things done by individuals, and thus attributable specifically to them. From this psychological frame how could we make sense of this fudging behaviour? At best the fudges can be interpreted as logical or psychological slips propped up by delusions of grandeur. At worst they represent academic chicanery and political manipulation in high degree (Nairn, 1981, p58).

If we regard this in a sociological context, however, a different picture emerges; psychometricians may well be regarded as the moral guardians of the age of competency, the high priests who hold society stable by propagating, preaching, and propping up the gospel of the Standard, and the cult of the linearly determined individual that it constructs and supports.

In the beginning

"What's in a name?" Bill Shakespeare said, "that which we call rose by any other name would smell as sweet." Maybe so, yet that which we call a trait when it is just a mathematical function takes on a different odour indeed. Names have a magic of their own, and the stickiness of the name is very dependent on the power of the namer.

Lord (1980) produced the seminal work on item response theory, in his book Applications of item response theory to practical testing problems. It is possible here to trace in detail the birth of a fudge.

Early on there are some laudably honest statements:

> True score theory shows that a person may receive a very low test score either because his true score is low or because his error score is low (he was unlucky) or both (p5).
>
> The true score is a mathematical abstraction. A statistician . . . does not try to define the model parameters as if they actually existed in the real world. A statistical model is chosen, expressed in mathematical terms undefined in the real world. The question of whether the real world corresponds to the model is a separate question to be answered as best we can. It is neither necessary or appropriate to define a person's true score or other statistical parameter by real world operational procedures (p6).
>
> In item response theory . . . the expected value of the observed score is still called the true score (p7).

Admittedly, our laudability quotient diminishes as we reflect on the use of the word "true." In what sense can it be true if it doesn't exist in the real world? Why call it true if it can't be measured. But perhaps it is true in a mathematical sense because it is a necessary conclusion for the premises of the theory? Not so, it is merely the name of a variable assumed in the theory.

Undeterred we press onwards. Five pages later Lord commences the serious work in developing the theory:

> Let us denote by ø the trait (ability, skill, etc) to be measured. For a dichotomous item, the item response function is simply the probability Pø of a correct response to the item. . . it is very reasonably assumed that P increases as ø increases (p12).

Now this is truly remarkable paragraph. The word "trait" has not appeared before. Where did this "trait", this "ability", this "skill" come from that is being measured? What does it mean? Lord "very reasonably assumes" that as this thing increases, the probability of answering a particular test item increases. But why do we need this thing at all? And why is it named a trait or a skill or an ability, which are hardly "mathematical parameters"?

We wait expectantly till page 45 to find out what ø means mathematically. "A person's number right score . . on a test is defined . . . as the expectation of his observed score x. It follows immediately . . that every person at ability level ø has the same number right true score." Then on page 46 the crucial point finally emerges "true score . . . and ability . . . are the same thing expressed on different scales of measurement. " And just in case you missed it, the best estimate of this true score, this ability, is the number of items answered correctly on the test.

Thus on his own admission Lord has done exactly what he claims statisticians do not do. He defines the parameter as having "real world" status when he calls it ability. (Just as he infers it has some objective or propositional reality when he calls it true). Its mathematical status is simply the number of items answered correctly under the idealised conditions specified in the theory. It's empirical status is the actual number of items answered correctly, or some statistical manipulation of that number.

There is one more aspect of this fudge that we need to look into. It is the fascinating use of the adjective "latent" in front of trait. Hambleton & Swaminathan (1982) elucidate:

> Any theory of item responses supposes that, in testing situations, examinee performance on a test can be predicted (or explained) be defining examinee characteristics, referred to as traits, or abilities: estimating scores for examinees on these traits (called 'ability scores'); and using the scores to predict or explain item and test performance. . . Since traits are not directly measurable, they are referred to as latent traits or abilities. Any item response theory specifies a relation between the observable test performance and the unobservable traits or abilities assumed to underlie performance on the test (p9).

Of course, this is not quite true. Item response theory does nothing of the kind. It assumes certain characteristics of test items, and then generates a total score which is an estimate of the true score. Under certain conditions, "we can think of ø as the common factor of the items" (Lord, 1980, p19). The true score can only be guessed. The mathematical theory tells us the probability that it lies somewhere within a certain range of scores. Latent means hidden or concealed or potential. What is hidden, what is latent, is not any characteristic of the person, but a characteristic of the measurement itself. The examinee has performed, has participated in the event of answering test items. Nothing hidden or latent about that. So why the displacement? How did a latent measure become a latent trait?

Item response theory doesn't need any assumption about traits at all. The talk of traits and abilities is redundant and gratuitous. After all the terribly refined and elegant statistical manipulations, Item response theory simply produces a total score which (given knowledge of the structural characteristics of individual items) allows a prediction of the probability with which any particular item will be answered correctly by a person with that total score. It does require a certain consistency of correct (or incorrect) response for specific items on the part of the examinee. All else, as far as item response theory is concerned, is fantasy.

Incidentally, such prediction is in no way an explanation; to assume that is to evoke the dormative principle; the total score is just a summary of information about a particular person answering the individual items. Such a score cannot now be used to explain why the items were answered correctly.

On page 55 Hambleton and Swaminathan (1982) come clean; rather by accident that design, I fear. "Ability", we read, "is the label that is used to describe what it is that the set of test questions measures." Precisely. And what it measures is an estimate of probabilities of answering certain test items correctly. To what extent that measure relates to any "characteristic" or "trait" or "ability" of the examinee may only be known after "construct validation studies . . . (which) validate the desired interpretations of the ability scores" (p55). Shouldn't that read "validate or invalidate"?

Mistakes: probability, correctness, and checking

Item response theory cannot predict whether a particular person (whose true score we don't know but whose estimated score we do know), will get a particular item (whose characteristics we know), correct or incorrect. The theory will predict the probability of getting it correct. In practice it will either be correct or incorrect (probabilities are only 1 or 0).

So item response theory never even pretends to estimate what people know or can do. It only claims to estimate the probability that they can do certain things. Then the assumption (and that's exactly what it is) is made that this indicates an ability of the person in that area of cognition. It might mean something else. Or it might not.

When I worked as a test constructor I noticed one aspect of answering tests that was interesting. When groups of year 10 students did the 100 item tests most would finish in about ninety minutes. When groups of year 8 students did the tests most would finish in about 60 minutes. The year 10 students got slightly better results (about 0.3 S.D. better). Conventionally this would be interpreted as meaning that they had more ability, or simply more maturation. But given my perceptual data, perhaps it just means that they did more checking!

Psychometric selection myths and fudges

Hulin, Drasgow & Parsons (1982) complain that the controversy and rhetoric about standardised educational admission tests seem to have developed independently of the psychometric evidence about the usefulness of admission tests in reducing errors in prediction. They claim that Cleary, Humpreys, Kendrick, & Wesman (1975), Rubin (1980), Linn, Harnisch, & Dunbar (1981) among others, have produced summaries of large numbers of studies relating college and professional school admission test scores to performance in post secondary and postgraduate educational institutional institutions:

> The evidence is clear and consistent. Well-constructed tests of cognitive abilities are significantly and consistently related to performance in school. When appropriate corrections are made for restriction of range and other statistical artefacts, the validities of tests are appreciably large (p 281).

Claims such as this are very common. So on this occasion I thought I'd check out the references.

Cleary's (1975) data involved correlations between verbal and mathematical SAT scores on the one hand and High School grade averages and College grade averages on the other. The correlations ranged from 0.35 to 0.50. But the correlations between the High School and College grades were higher at 0.64. So two points about Cleary's study: firstly the correlations are at best only 25% better than pure chance. Is this "appreciably large"? Secondly, they were considerably lower than the correlations from grade averages, so why were they necessary at all?

Rubin's (1980) study involved the use of the Law School Admissions test to predict first year grades in 82 law schools. The correlations ranged from 0.03 to 0.5; after corrections for range (Linn, 1981), the correlations range from 0.2 to 0.7. In 14 of the schools they were below 0.35, which is 12% better than chance. Is this "appreciably large"?

When it is known that issues of construct validity introduce far more sources of error than are involved in simple predictive correlations of this sort, it is difficult to understand how this sort of justification, which is quite common in the literature, goes on for decades virtually unchallenged within the psychometric community; on the other hand, compared to the abysmally low correlations often obtained in such predictive correlational studies, perhaps they are appreciably large.

However, these studies raise another issue and another fudge; the correction (always upwards) of predictive correlations.

Fudging the predictive correlations

Correlations between a selection instrument and later performance are often corrected for range restrictions and for criterion unreliability. Range restriction is reasonable; generally some of the people tested were not selected, so had no opportunity to be in the final sample. It is considered appropriate by statisticians then to estimate what the correlation would have been had all of those selected actually been appointed. After the correction, of course, it is a correlation about something different; it becomes the estimated correlation between test performance and later performance of all those who sat for the test. Prior to the correction it was the correlation between test performance and later performance of all those who performed later. Different sample, different correlation. Which to use depends on what question you ask. Automatically raising the correlations is a fudge.

Correcting for criterion unreliability is a different matter. Most job tasks are multi-dimensional; that is, they involve many very lowly correlated tasks. And college grades are likewise composites based on lowly correlated components. If a single correlation is to be obtained a with multi-dimensional job performance the various ranks or gradings have to be collapsed into one single rank or grading; and that requires some arbitrary and explicit loading to be applied to each dimension. (See Chapter 10 on Comparability).

Even when this is done (and it often isn't), there is still the assumption that there is indeed a meaningful rank order to be obtained. If most people in most jobs or in most courses do their work adequately (just as most people drive cars adequately), then we would expect correlations to be low, and ultimately, where training schemes are very adequate, to be zero. In such situations, the reliabilities would be low not because of rater inadequacy that can be corrected for, but because raters are attempting to separate performance when it cannot be separated, or/and are trying to pretend that a multi-dimensional performance is in fact uni-dimensional. In such cases it is obviously not appropriate to artificially inflate the correlations because of rater unreliability.

The changes are more than trivial. A study by Schmidt, Hunter & Pearlman (1981) involved 150 000 people, 2000 predictive correlations. Before correction the average correlations between eight aptitude tests and job performances in clerical job categories ranged between 0.15 and 0.25. After the statistical corrections, however, they magically rise to between 0.3 and 0.5. Still not good. In fact, still quite awful. But they certainly look better than before, and aptitude tests survive again to live another day.

The great Queensland reliability fudge

I was talking to the Principal of a secondary school in Queensland. Students in year 12 are assessed internally, with the help of some external monitoring. I suggested that there might be some problem with reliability. "It's 0.95," he replied with confidence. "Excellent," I responded with some scepticism. Then I decided to check the data.

The study is titled Random sampling of student folios: a pilot study (Travers, 1994). In this study

> . . . 1189 exit review folders of Year 12 student work were collected randomly from school subject groups across Queensland in December 1993 and assigned to two hundred and forty review panellists in other districts. These exit review folders show the work of students who have received a result for that subject on their Senior Certificate. The role of the review panellists was to examine packages from schools containing ten folios, and for each folio decide a Level of Achievement and relative position in that achievement band (p 1).

The review panellists were given access to other marker's assessments and comments, as well as the school's assessment of the Level of Achievement. What they didn't have was information about the rung placements within each level of achievement (There are ten rung placements within each level of achievement) .

So this is not a blind reliability study:

> because it was not possible to reproduce all the conditions under which judgments about students were made by schools which supplied folios. In particular, panellists did not have the opportunity to observe student performance over an extended period of time as teachers do (Travers, 1994, p12).

The astute reader will already have noticed a contradiction here. The study was not constructed as a blind reliability study where no previous marks or grades were attached because they wouldn't have sufficient data to make valid judgments about levels of achievement. On the other hand they are being asked to make much finer discriminations regarding rung placements.

The astute reader will also doubtless have expected a very large halo effect, and would not be surprised if reliability coefficients, at least in relation to levels of achievement, were very high. As indeed they were. Eighty per cent of achievement levels remained unchanged, most of the aberrant cases being one level lower, indicating, no doubt, the "high standards" of the review panellists.

The overall correlation figure obtained for agreement between school exit and review level rung placements, on a fifty point scale, was 0.95. The authors were particularly pleased with the rung placement data:

> a rung difference of plus or minus one or two is not so much a significant difference as a demonstration of precision and accuracy . . . half the decisions about rung placement involved either assigning the same rung or one or two rungs lower. . . (this) suggests that not only do these panels read the folios very closely, but that they are able to arrive at decisions about standards that are both highly reliable and very precise (Travers, 1994, p17).

I did a little experiment. I listed fifty (hypothetical) folios in rank order of one to fifty, with ten papers at each level of achievement. Then, keeping them at the same level of achievement, randomly allocated new (reviewed) rung placements within each level. The rank order correlation was 0.95.

It follows that acceptance of given levels of achievement (halo effect), combined with random allocation of rung placements, is sufficient to account for the 0.95 correlation that was used to justify the whole procedure, not only of the pilot study, but indeed for the whole examination system, as evidenced by the Principal's comments.

Rather than evidence of precision in rung placements, which determine tertiary entrance scores, the data generates evidence of randomness, and another psychometric fudge is perpetrated by well meaning psychometricians on a gullible public.

The General frame and the true score

The General frame of reference as hijacked by psychometricians contains as an essential element of its assumptions the notion of a true score; a further element of those assumptions contains the notion that it is possible in some way or another to approach that true score; to get measures empirically closer to the true score by various procedures implied by the particular model. For example, in classical test theory by increasing the number of items on the test; in generalisability theory by sampling more tasks more randomly from a bigger collection of possibilities; in item response theory by having more items of appropriate characteristics which are uni-dimensional; in domain referenced tests by having the domain of items criterion referenced to a high degree.

Allied to this frame but not tied to it so tightly are the various notions of reliability and validity that have not been developed as part of the mathematical models mentioned in the previous paragraph, but have emerged from more general considerations of the notions of assessment, rather than of tests. In my terminology, these considerations have challenged the artificial constriction of the general frame by psychometricians, and have restored, through notions of construct validity and consequential validity, at least some of error components previously bypassed.

However, this has produced a contradiction with the notion of the true score that has not been made overt. For example, as described in Chapter 16, most achievement tests are not made more valid by increasing their reliability; on the contrary high reliability is seen to be, in most circumstances, an indicator of low validity. For most achievement areas involve a large number of disparate activities, and there is no a-priori, or even post empirical reason to believe that these activities are uni-dimensional, or otherwise closely inter-correlated.

I argue in Chapter 15 generalising the assessment events across contexts, or time, or media, or even value assumptions or frames of reference, does not (as does generalising across selection of test items or markers), reduce the standard error of the estimate; on the contrary, we have every reason to believe that it will increase such error, to a point where the whole notion of true score becomes unsustainable. After all it is not by chance that so much space is given in test manuals to ensuring the conditions under which the test is given are kept constant. Obviously this indicates the fragility of the test to contextual shifts. (On second thoughts, it could be as much a ritual designed to imply scientific accuracy, and sustain the notion of fairness). Regardless, it is clear that contextual shifts increase the error term, whilst contextual control artificially reduces it; artificially because no argument is ever given, nor could it be sustained, that this particular test context is superior to any other to the measurement of this "ability." So once again the price of higher reliability is lower validity.

Preview

We could go on dealing with the specifics, but it is time to present the greatest fudge of all. Validity. For as will become clear, the very definition of validity creates a discourse around it where every test may be assumed valid until proved otherwise, and as there are no specific descriptions as to how such a proof might be constructed, and no specific standards of acceptability to which such descriptions might be compared, all assessments may claim to be valid.

Return to Table of Contents

Chapter 16: Validity and Reliability

Preview

The professional theoretical face of assessment discourse asks the question, is the test reliable? More ethically orientated assessors ask the additional question, is the assessment valid?

The public wants to know, is it fair? And the more critical of them might add, are people being violated?

In this chapter some of the more recent work on validity is discussed, and its positioning as advocacy demonstrated.

Reliability is also discussed as a problematic, rather than as an obvious prerequisite to validity.

Validity

"Validity," states the first sentence of the APA Standards of educational and psychological testing (American Educational Research Association, 1985), "is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (p9). It goes on immediately to explain that: "Test validation is the process of accumulating evidence to support such inferences."

Which all sounds very scientific and objective and devoid of bias. But is it so? Let me, from my own particular concern with the test taker, rewrite the first sentence to dovetail more accurately with my concerns.

"Invalidity," states the first sentence of the alternative tract, "is the most important consideration in test evaluation. The concept refers to the inappropriateness, meaninglessness, and uselessness of the specific inferences made from test scores. Invalidity or error estimation is the process of accumulating evidence to problematise and ultimately reject such inferences."

It should be clear even from this small rewrite that a text that began with the second conceptualisation would be a very different text from one that began with the first.

Positioning

The main participants in the testing process, we are told, are the test developer, the test user, and the test taker. Also often involved are the test sponsor, the test administrator and the test reviewer. Sometimes, many of these participants may be parts of the same organisation, with the notable exception, of course, of the test taker.

As clearly stated in Chapter 1, my position of value, my backdrop when I seek information about events, concerns the violations perpetrated on the participants in those events. So in the matter of testing, my focus is on the test taker, and in what ways the taking of tests and the inferences and consequences flowing from such events constitute a violation - a diminishing of personhood, a misrepresentation of potential or action, a claim to unwarranted accuracy of description, and thus unwarranted control and construction of the living human person who is taking the test.

The 1985 Standards acknowledge, with fine understatement, that "the interests of the various parties in the testing process are usually, but not always, congruent" (p1). This trivialisation of the traumatic effects, dislocations, and exclusions of millions of students based on test and examination results is quite remarkable. Perhaps it is just another example of the way social positioning can overwhelm interpersonal sensitivity and intellectual honesty.

The concern of the test makers and users is, after all, with hundreds, thousands, or hundreds of thousands of test takers (not to mention their concern with their Board of Directors and shareholders). But their concern is with them, viewed as a group. Their interest is with groups, not individuals; in summaries, not raw data; with simplifying complexities, not with complexifying individuals; with objectifying human subjects, not with subjectifying human events.

For the test constructor, sponsor and user there are so many difficult questions; so many criteria to consider; so many factors to consider if the overt and covert claims of the test makers are to be defended. We shall deal with these in due course. Yet to the test taker there is only one question, a normative question which emerges from his or her very construction as an individual. Have I passed or have I failed? Am I satisfactory or unsatisfactory? Am I normal or a nut case?

Additionally and ironically, it is precisely because they see the testing event from this individualised perspective, rather than from a group perspective, that they do not ask the more crucial, the more fundamental question: How much error, ambiguity, uncertainty, does this attribution contain? Or is it their powerlessness, and unheard voice, that makes these questions at the best unspeakable, at the worst unthinkable?

Sources of evidence

The 1985 Guidelines describes an ideal validation as including

> several types of evidence. . . Other things being equal, more sources of evidence are better than fewer. However, the quality of the evidence is of primary importance, and a single line of solid evidence is preferable to numerous lines of evidence of questionable validity (p9).

This is hardly reassuring for the test taker. The tautology and redundancy in the phrase "questionable validity" is remarkably inept; validity is proposed as the characteristic of the evidence used to support the construct "validity," and the essence of the concept is surely its very questionability. Far more damning, however, is the clear implication that evidence that does not cogently support the assertions of the test users should not be presented. Putting it another way, validity is a concept based on advocacy, is a rationalizating tool for a methodological decision already made, and is an ideological support rather than a scientific enterprise.

Is this an over-statement? Here is the first sentence of the next paragraph of the 1985 Standards: "Resources should be invested in obtaining a combination of evidence that optimally reflects the value of a test for an intended purpose" (p9). The word "optimally" says it all.

So, validity is clearly an advocacy construct, based on the assumption that any assessment data is innocent until proved guilty. The discourse about validity presents the case for the defence. There is no advocate for the prosecution, so the prosecution case does not present its case. More than this; the very idea of a prosecution case is denied by the definition of validity.

Yet here we also see, in the very heartland of post-positivist empiricism, the embryo of a discursive construct; an appeal, not to numbers, but to discourse. Over the next ten years Cronbach (1988) and Messick (1989a, 1989b,1994), doyens of psychometrics, in their born-again personas will enlarge the idea of construct validity to a point where Cherryholmes (1988) will nail it as fully discursive, and thus "linguistically, politically, economically, socially, culturally and professionally relative"(p450).

Even so, the advocacy position remains essentially unchanged. Messick(1989b) asserts that :

> To validate an interpretive inference is to ascertain the extent to which multiple lines of evidence are consonant with the inference, while establishing that alternative inferences are less well supported. This represents the fundamental principle that both convergent and discrimanant evidence are required in test validation (p1).

But note the implication of "are less well supported" and its relationship to advocacy. And later in the same article, when he gets specific about invalidity implications of adverse social consequences, he says:

> If the adverse social consequences are empirically traceable to sources of test invalidity, . . . then the validity of test use is jeopardized. . . If the social consequences cannot be so traced - or if the validation process can discount sources of test invalidity as the likely determinants, or at least render them less plausible - then the validity of the test use is not overturned (p11).

Note the use of the words "jeopardised," "less plausible," and "not overturned." Given the probabilistic nature of all social research, the chances of any test being declared invalid on the basis of these criteria, from this perspective, are remote. Ultimately, Messick is eminently logical. For if "validity always refers to the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores"(Messick, 1989a, p13), then even infinitesimal support, being support, makes the test valid, and nothing has really changed since Guilford's (1946) claim that "in a very real sense, a test is valid for anything with which it correlates " (p429). And as error will ensure that no tests correlate zero with anything, it follows that all tests are valid.

Reliability

Even though validity has taken on a post-modernist hue of recent times, reliability has, until recently, remained untouched as a "foundational" cornerstone of educational measurement. Reliability was seen as the lower limit of validity. An assessment could not be more valid than it was reliable.

The assessment industry, whether local, corporate, government, or quango, has embraced the reliability concept both ideologically and empirically. In contrast to validity, estimates of reliability are often obtained and circulated. There are two reasons for this: the reliability of the test can be measured using only data from the test scores; and often relatively high values (correlations of 0.7 - 0.9) can be obtained, if for no other reason that they are so constructed to ensure that such high internal consistency occurs.

Politically such reliability data can be used to "prove" the quality of the test, and maintain the illusion that reliability refers to "the degree to which test scores are free from errors of measurement," which is how they are described in the first sentence about Reliability in the 1985 Standards. In fact, the Standards emphatically insist that:

> For each total score, sub score, or combination of scores that is reported, estimates of relevant reliabilities and standard errors of measurement should be provided in adequate detail to enable the test user to judge whether scores are sufficiently accurate for the intended use of the test (p20).

Note that it is never suggested that the standard errors of measurement information should be available to test takers. There is a later chapter in the 1985 Standards entitled "Protecting the rights of Test Takers." Again there is not the vaguest suggestion here that such information should be made available to them.

However, even reliability is now under threat. Is there nothing sacred? Moss (1994), has cogently argued that there can be validity without reliability. She points out that:

> Reliability, as it is typically defined and operationalized . . . privileges standardised forms of assessment. By considering hermeneutic alternatives for serving the important epistemological and ethical purposes that reliability serves, we expand the range of viable high-stakes assessment practices to include those that honor the purposes that students bring to their work and the contextualized judgments of teachers (p5).

Such idiosyncratic behaviours and judgments tend towards a diversity that reliability abhors. There are two issues here. The first relates to the relationship between reliability and validity perceived from the standpoint of the assessors; the second deals with the concept of reliability, that is consistency, of performance as actually produced by the persons being assessed. The two issues are related in that they both relate to responses to persons involved in an event designed to describe what a person can do by asking them to do something else, and then making inferences about what they might do in another time and place and context.

Let's first look at this expectation of high reliability, and the theorising that precedes it. The argument is essentially this - if one test or examination is reliable then another similar test or examination will give the same verdict, however that verdict is communicated - as marks, grades, pass-fail, selected, or whatever. It is logical to assume, therefore, that one half of the test would give the same verdict as the other half, because all of the bits of the test contribute to the final score and hence the final verdict; putting it another way, we are dealing with some linear dimension here, some unitary idea or construct; all of the questions measure it with considerable error, but the more interconnected questions we ask, and the more inter-correlated answers we get, the more the error is reduced, and the more the measurement is refined to approach the true measure of it. Of one thing we are sure. The "it" is out there, waiting to be measured And "it" has a true value, that we can approach but never completely determine. This simplistic positivism is at the epistemological and ontological heart of educational measurement.

Teachers and public examination boards do not believe that this is what they are doing, even though the latter have no hesitation is using measurement theory to manipulate their results and rationalise their processes. They do not necessarily believe there is some unilateral trait or ability or skill that underlies the total score or grade. Indeed, as Willmott and Nuttall (1975) point out:

> In the field of 16+ examining it is quite possible that any increase in reliability would be to the detriment of validity. This is easily seen to be the case, since by refining questions and components so that they correlate highly is to learn more and more about less and less: the trait being measured is defined even more narrowly as reliability (in the sense of internal consistency) is increased. In such a situation, the validity of the examination concerned is bound to decrease owing to the narrowness of the field covered.

> A glance at any subject syllabus published by a CSE or GCE board shows clearly that the comprehension of a very wide variety of content is required of candidates and, in many cases, the educational objectives required of candidates in following the course are equally varied (p55).

It is a pity that these authors do not take this argument to its logical conclusion: that there is no single trait to be measured, that there is no linear concept to be categorised, and that there is no necessary correlation - indeed there may be some negative correlations, between the relative performances of candidates on various objectives. But this conclusion would lead inevitably to the final one, that there can be no meaningful rank order of students, because the rank order can give no meaningful information about the performance of individual students in relation to any particular objective (See Chapter 10 for a far more detailed description of the comparability issues involved).

Perhaps one more very simple example of this may be pertinent. Imagine a course in electrical wiring which has only two objectives; one relates to the safety requirements, the other to the ability to problem solve in practical situations. An examination is devised to measure the attainment on the course; half of the marks in the examination relate to safety requirements, and half to problem solving. Two students each obtain fifty per cent of the marks. What do we know about their attainment of the objectives? Nothing! One student may have got all the safety questions correct, and the other all the problem solving questions correct. In this case between them they may be considered to know everything, or nothing! In regard to validity, to inferences about objectives made from test scores, the validity has to be zero, if we focus on these individual students.

Note that the above argument is valid regardless of the correlations between the scores on the two parts of the paper for a group of students.

It can be seen that the reliability of the test in this case is irrelevant, as is any estimate of inferences that may be made about the group of students. For the group we could indeed make inferences about the probability that they knew, on average, a certain proportion of the safety information, and could solve a certain proportion of the problems. But just as a total score loses all the information about individual questions, so does it lose in this case all the information about individual students.

Incidentally, correlations across different subjects are often also of the order of 0.8. That is the correlation between two tests of different subjects is about as high as the reliability of any one test. (quoted by Nuttall & Willmott, 1975, p48). Perhaps there is a linear trait after all, but unrelated to the apparent construct being measured. What might this construct be? Traditionalists would be in no doubt that it was a general ability that they would label intelligence. Yet we know that the correlations between examination scores and other sorts of measures (eg, job performance) are very low, of the order of 0.3. So a more direct and sustainable interpretation is that "it" is the ability to perform in the events constructed around examinations. Examinations measure examination ability!

The second issue is rarely mentioned in the literature, and it relates to individual consistency of performance. An example might be taken from cricket. Batsmen vary in the consistency of their performance. Consider two batsmen who each has an average of about thirty runs over a large number of innings. One may score very consistently between 20 and 40 runs. Another may score the odd century, but may often make less than 5 runs. Test theory cannot account for this. It defines 30 as an approximation to their "true score," the score that best matches their "batting ability." But any deviation in a particular innings would be attributed to "random error," and be expected to assume a random rather than a consistent pattern. What becomes obvious from this example is that the average (true) score for these two batsman has a very different meaning; while for one it may indeed indicate the "most likely" score, for the other is indicates a most unlikely score indeed.

A fundamental contradiction

Now this argument, if we take it a little further, leads to a very strange conclusion. Let's go back to the first line of the Willmott and Nuttall (1975) quote: "it is quite possible that any increase in the reliability would be to the detriment of validity"(p55). They show why this is so in the measurement of any multi-dimensional area, and Moss (1994) indicates why it is so for "hermeneutical alternatives." But increase in reliability from what point? From 0.8, or from 0.5 ? Or from zero? Is there an argument to be made that all reliability negates validity. This would lead us to the apparently absurd conclusion that the greater the reliability the lower the validity, and the ultimately maximum validity is to be obtained from zero reliability. In terms of measurement, this would mean, of course, that human "constructs" were essentially unmeasurable. We can talk about them, but we can't measure them. Which is what Cherryholmes (1994) is really saying when he says the "construct validity is fully discursive." Isn't he?

In the next chapter I list thirteen sources of error, thirteen sources of invalidity. Two of these, related to multi-dimensionality and values, are dealt with by Willmott and Nuttall, and by Moss. What of some of the others? Do they show the same pattern of an increase in reliability leading to a decrease in validity?

Temporal errors are certainly increased by calculating reliability on the basis of one test at one time. As performance would be expected to vary with occasion and over time, one shot assessment certainly decreases validity error as it increases reliability

Contextual errors are certainly increased by confining assessment to pencil and paper situations and producing a very singular and artificial environment in which the assessment occurs, to the extent of standardising format and time available to complete the tasks. Again reliability is obtained at the expense of validity, which implies generalising to other contexts.

Construct errors are likewise increased through the limitations of content, form, process and media that is determined and narrowed through the testing or examination procedures. Again the capacity to generalise, and thus the validity, is diminished by the psychometric strictures required for high reliability.

The effect of high reliability on categorisation errors is complex. Where categorisation is defined in terms of percentiles of the group tested, categorisation errors are reduced as reliability increases, leading to an increase in validity. However, when one particular marking scheme (rather than another marking scheme) is used to increase the reliability, the reduction in categorisation error is illusory rather than real. And where comparability issues intrude, meaning fogs up as psychometric solutions compound the categorisation problems. So in these areas the effects of reliability on validity are moot.

In similar vein, errors attributable to frame of reference shifts, to labelling and attachment confusions, to prediction inaccuracies, or to logical type confusions, are largely indifferent to reliability. And whilst consequential errors, the negative effects of testing, have certainly been exacerbated by the quest for higher reliability, it is the quest rather than the empirical value that is involved.

Instrumental errors of course are reduced as reliability increases; indeed, reliability may be defined as the inverse of instrument error. So in this one area it is clear that increases in validity are dependent on increases on reliability. Yet if, as we have shown, the effect elsewhere is that such increase in reliability either decreases validity or has an indeterminate effect on it, then the general proposition holds, and we may say that in the empirical world, the procedures used to increase reliability result in a decrease in validity.

Born again validity

Messick (1989a) has broadened the concept of validity to refer to "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores"(p13), and this includes the way values "influence in more subtle and insidious ways the meanings and implications attributed to test scores"(p59), so that "test validation embraces all of the experimental, statistical and philosophical means by which hypotheses and scientific theories are evaluated" (p14).

Messick's position seems to be generally accepted. The sources of potential error actually referred to do cover the range and depth of epistemological, ontological, and value sources referred to in this thesis. Yet even with this multiplicity of error, this proliferation of possibility of miscategorization, Messick (1989) insists that validity is a unitary concept, a singular "degree of support":

> The essence of unified validity is that the appropriateness, meaningfulness, and usefulness of score-based inferences are inseparable and that the unifying force behind this integration is the trustworthiness of empirically grounded score interpretation, that is, construct validity(p5).

In other words, validity is a statement of faith in testing, a statement of justification by an "expert" that the whole assessment event is legitimate, is valid. Even though, in practice, for real tests, the considerations and scientific inquiries that Messick advocates are rarely carried out.

Let's look at this in more detail; first it is apparent that appropriateness, meaningfulness, the usefullness are sometimes quite separable. Appropriateness applies very much to particular values. In my value system, any test which violates individual students is inappropriate. Yet it might be quite meaningful in that some inferences made from it can be understood and acted on by teachers and administrators, and it may be useful in that predictions made from it help selection processes. In another case a test of inverted neuroticism may be quite useful in predicting successful medical students, but may be considered inappropriate for that application. It's meaningfulness may be moot. Ultimately, of course, the very meanings of appropriate, meaningful and useful are deferred; they are partial synonyms for valid, the word they supposedly elucidate.

It becomes clear that the "unifying force" then is not created by the congruencies among appropriateness, meaningfulness and usefulness, but rather by the "trustworthiness" of the "interpretation." In other words, by the power that resides in the status of the "expert" who controls the discourse in which the judgement is embedded. And because the discourse of validity is in essence about all the ways in which the measurement cannot do all the things it claims to do, and explicitly about some of the ways it might be done better, an advocacy judgment would concentrate on some way or ways in which the test was better than it might have been had such improvements not been made. According to Messick, this is the unifying force that asserts, and thus proves, validity.

Specifically, my analysis of Messick's (1989a) definitive paper in the third edition of Educational measurement indicates that he makes reference to over fifty sources of potential invalidity; for indeed, how can he describe how a test may be valid without focussing on all of the ways in which it might not be valid. I have indicated some of these references, and their relation to the error sources that I specify, in the next chapter.

Finally, the very existence of validity is established, validity is indeed made manifest, through the denseness of the arguments used to refute such existence, together with the reassurance that the battle continues, and some gains have been made.

Let me be specific: The definition of the construct of validity does not exclude the notion of invalidity. However, the discourse on validity, constructed as it is from the position of advocacy, excludes the notion of invalidity as an issue. More than this, the discourse itself becomes the arbiter of the proof of validity claims, independently of empirical data, which becomes irrelevant within the density and complexity of the discourse; as a result, empirical data to justify validity claims is rarely collected, and when it is it is inevitably construed as supporting the claim. Evidence rejecting the validity claim is never collected because such positioning is absent from the discourse. Madaus (1986) puts it nicely:

> present methods of gathering content validity evidence are inadequate; they are designed in such a way as to almost guarantee a positive outcome. Alternative methods designed to disconfirm or test counter hypotheses about the issues are, in my experience, never employed (p12).

Practically, the psychometric scam is accomplished by focussing on the test score, and ignoring its dark side, the standard error of estimate; specifically, by implying that the estimated score is the true score, that the intention is the empirical fact, that talking about problems of validity magically increases validity, and that increasing validity makes a test valid.

Validity and the predominant paradigm

When advocacy is positioned, aligned to the predominant paradigm, then advocacy is interpreted as truth. Truth not as the production of true utterances, but in Foucault's (1982) sense of "the establishment of domains in which the practice of true and false can be made at once ordered and pertinent"(p8). From the 1980s, when the prevailing societal metaphor is the discourse that surrounds economic rationalism, and in particular those myths connected with people competencies, the metaphor is rabidly post-positivist, and validity definitions (advocacies) based on those assumptions will be seen as self-evidently true. As Cherryholmes (1988) puts it from his post-modern perspective: "boundaries limiting construct-validity discourse have yet to be justified. They are policed nonetheless "(p154).

In contradistinction, advocacies for more post-modern descriptions (eg validity characteristics for qualitative research) are clearly not aligned to the prevailing world-view, and so will be interpreted as justifications. They advocate from a loser's position, so at the best their views are accepted as tentative, at the worst as unproven and hence unacceptable assumptions. This is inevitable because no abstraction can be proven to be correct, so acceptance is always a function of value, rather than of rational proof; and moral value is usually construed as stabilisation of the status-quo, as confirmation of the predominant paradigm.

Shepard (1991) gives an example: "measurement specialists asserted that performance assessments are less reliable and less valid than traditional tests and that they are potentially biased because they rely on fewer tasks." But then she adds: "Why are existing tests presumed to have the high ground in this dispute? What claim do traditional tests have to validity?" (p10).

This is not to deny the acceptance of such advocacy in localised communities (eg some faculties of some Universities) where a paradigm shift has already occurred.

Qualitative assessment and qualitative research

Validity criteria in qualitative assessment has lagged behind validity in quantitative research. However, the two fields are closely aligned. In fact Messick (1989a) regards then as virtually synonymous in that

> test validation in essence is scientific inquiry into score meaning - nothing more, but also nothing less. All of the existing techniques of scientific inquiry, as well as those newly emerging, are fair game for developing convergent and discriminant arguments to buttress the construct interpretation of test scores" (p56).

I do not want to focus on the blatant advocacy aspects of this statement implicit in such terms as "fair game" and "buttress," but rather on its implication for using research validation criteria for assessment. In addition, I would want to include "categorisations" as a limiting aspect of "score." With this addition the work done on qualitative criteria for research validity becomes appropriate for assessment validity.

Summary

We have worked our way through some of the minefields of validity and reliability discourse. In particular I have indicated how the notion of advocacy built into the very definition of validity overwhelms scientific detachment, and effectively silences the logical inferences that derive from the voices of confusion and error that are the very basis of validity discourse.

The emphasis on reliability of assessment instruments is also shown to be a misplaced source of credibility for assessment, because measures to increase reliability are shown to decrease validity.

Now the coin can be flipped. The underside of validity can be examined. The nastiness of error can be exposed. In the next chapter the sources of invalidity are spelt out in detail.

Return to Table of Contents

Part 5: Synthesis

Chapter 17: Error and the reconceptualising of validity

Preview

From the analysis so far, it is possible to produce a general definition of error as it applies to the field of educational measurement and/or categorisation. This is the flip side of validity which exposes that general nastiness called invalidity.

In this chapter the notion of invalidity is reconceptualized, having both discursive and measurable components. Thirteen (overlapping) sources of error are examined, all contributing to the essential invalidity of categorisations of persons. For easy reference I have indicated the summary theoretical and practical definitions of these error sources in bold print.

Definition of error

Error is predicated on a notion of perfection; to allocate error is to imply what is without error; to know error it is necessary to determine what is true. And what is true is determined by what we define as true, theoretically by the assumptions of our epistemology, practically by the events and non-events, the discourses and silences, the world of surfaces and their interactions and interpretations; in short, the practices that permeate the field.

All assessment statements about a person are statements about that person engaged in an event, or a potential event. They are descriptions or indicators or inferences about the person's performance in that event. As such they involve at the very least an event in which the person being assessed is an element, and an event in which the assessor engages directly in the first event, or with a product (element) of it.

Error is the uncertainty dimension of the statement; error is the band within which chaos reigns, in which anything can happen. Error comprises all of those eventful circumstances which make the assessment statement less than perfectly precise, the measure less than perfectly accurate, the rank order less than perfectly stable, the standard and its measurement less than absolute, and the communication of its truth less than impeccable.

I want to list some of those sources of error, some of the conditions that change the measurement of a standard from a thin red line into a broad blue band: In doing so I will reject the notion of construct validity as a unitary concept, and dismember its dark side into disparate if sometimes overlapping categories.

Sources of error

I have named these sources of error:

1. Temporal errors
2. Contextual errors
3. Construction errors
4. Labelling errors
5. Attachment errors
6. Frame of reference errors
7. Instrument errors
8. Categorisation errors
9. Comparability errors
10. Prediction errors
11. Logical type errors
12. Value errors
13. Consequential errors

1. Temporal errors

We would hope our description of performance would have some substance; would be a stable quantity, invariant over time and space, rather than some ephemeral numerical butterfly attaching itself momentarily to the person assessed. If the person's performance is described differently if done at another time, in another place, with another group of people, then such difference as there is represents a source of error.

Or is it? Should we rather discount stability as being counterproductive in an educational situation? If stability is seen as the very antithesis of the educational enterprise, which we could define as being dedicated to change, then we would not wish any description to remain stable, as this would represent a nullification of the educational process.

Contrarily, if we wish to maintain stability as a criteria for assessment accuracy, we must be certain that all learning pertaining to the performance ceases at the time of assessment. And that none occurs during the assessment process. As well as all forgetting for that matter. Otherwise the error of the description increases rapidly, as the permanency of the description becomes increasingly dismembered by the ravages of time.

Regardless of which side of the fence we want to sit, or whether we want to sit on the fence, pretend it isn't there, and attribute the concomitant pain to other variables, stability must logically remain as a pertinent, or in conventional circles an impertinent criteria, to be considered in any estimate of error in assessment. My conclusion is that the logic of its contradictions makes most of the academic and psychometric definitions of reliability trivial.

So temporal errors have their genesis in changes that occur over time; persons change over time; tests change over time; the "same" event has different meanings over time. People are not computers, they react differently at different times; and they forget. So temporal errors increase over time. (Not to mention that different people make different meanings out of the same event; which makes it, of course, a different event.)

Temporal errors thus include all those confusions that constitute the dark side of stability, one aspect of reliability.

Practically, temporal errors are indicated by the differences in assessment description when the assessment occurs at different times

2. Contextual errors

Contextual errors constitute the underside of claims to generality and generalisability.

Any performance is relatively specific and defined: It is a single instance of possible instances; it is an event chosen from a multitude of possible events; it is a particular designed to illustrate a generality. Yet the performance will invariably be described (labelled) in terms of the generality it aspires to, rather than the specifics that define it. This is true of almost any evaluation, any test that goes beyond the description of a single behavioural objective, and even that, one step back, will often be found to be illustrative of a class of objectives, rather than of particular significance in its own right.

In the old days (good or bad depending on our values), this would constitute an example of "transfer of training." The claim was that if you could think clearly in Latin, then this should transfer to dealing adequately with the complexities of life in the social world; or if you could think logically in mathematics, then you could do so in international affairs; not to mention playing Rugby being a necessary prerequisite to running an Empire. When empirical data showed that such transfer was tenuous, the notion was kept, but the name changed. Taxonomic terms such as application and analysis, or the more up-market process called problem solving, have latterly laid claim to this temporarily non-habitable area. As well, the notion of a "skill" has latterly become fashionable, and generalisable social, cognitive, emotional, spiritual, and psychomotor skills proliferate, securely untrammelled by prophylactic empirical data of any kind.

As soon as assessment descriptions are committed to paper, their material permanency is dramatically increased. Likewise, the span of their associations is spread and emphasised. No longer just a description of a particular performance, the assessment becomes interpreted as a measure of knowledge and ability, an indicator of achievement on a course of study, and a predictor of future success or failure.

One source of error then is the magic transformation that occurs between numbers and categorisations, between specific acts and generalised descriptions. Unless the assessment statement purports to be no more than a statement about a particular assessment event, then the differences between this statement, and those obtained from all other possible contexts, is error; these are the generality differences attributable to other equally relevant contexts, eg written, oral, cooperative, on-the-job; all those boundaries that possibly could contain the assessment event that are different to the boundaries of the particular assessment event. Context also includes those power relations that pervade it and the judgment processes embedded in it that affect the performance of the person assessed, and the judgment of the person assessing; and this includes those that the boundary localises, as well as those that invade its permeable surface.

Contextual errors contain all the ambiguities inherent in those relations and elements and discourses that impinge on the event, but get excluded from the label.

Practically, contextual errors include all those differences in performance and its assessment that occur when the context of the assessment event changes.

3. Construction errors

The performance that is described in an assessment is generally built up of a number of parts; a science test is built up from a number of questions; an electrical automotive practical test requires the identification and repair of a selection of common electrical faults; a social skills assessment requires gradings on a number of interactional criteria, or more likely a game constructed about such criteria in multiple choice form. Such constructions are designed to represent the course of study, or the skill requirements, or the criterion referenced framework, that the assessment is supposed to describe. Further back still, the course has itself been constructed to improve performance in some areas of living, in some role as citizen, home maker, academic, engineer, baker, or whatever.

Somewhere, sometime, someone must make a choice about how far back along the chain of constructions we go in order to estimate the error, the difference between the "perfect" description of performance and the actual one that our assessment produces.

Let's take the electrical automotive test as an example. We could begin with a requirement to describe how well a student could identify and repair any electrical fault on any car brought into any garage (A). From this we construct a thirty hour course of study called Automotive Electrical Mechanics 2M, complete with course aims and objectives and assessment criteria (B). From this we construct a one hour pencil and paper test (C) and a two hour practical assessment (D).

Now how are we to describe the construction error in assessing a particular person? Is it the difference between the descriptions given in C and D? Or the difference between the matches of B and C on the one hand, and B and D on the other? Or should we look at the matching between C and D and A? Or is it all of these?

> *Why don't you describe A directly?*
>
> *You mean put people who've done the course into a garage and see how they perform?*
>
> *Yeah. Why don't you do that?*
>
> *It would be very expensive to do it for everyone?*
>
> *You don't have to do it for everyone. Just for enough people so you'd know if there was an error.*
>
> *There's always going to be an error.*
>
> *OK, so you find there's an error. If it was a small one then you could assume that the course, or the test at the end of the course, was well constructed because it did what it was supposed to do.*
>
> *That would be nice.*
>
> *And if there was a big discrepancy then you'd have to do it*

> *different.*
>
> *Do what different?*
>
> *I dunno. You're supposed to be the expert. Do the end-test different. Or do the course different. Or it might be easier to find another garage.*
>
> *Bit dangerous. There could be a lot of people get upset if we did that. No telling what sort of litigation we might run into if we found that the course didn't do what we said it would do.*
>
> *So ignorance is bliss, huh?*
>
> *Certainly not. We just need to be very careful, in terms of spending time and money on obtaining information that at the best will be useless, and at worst will only erode confidence and create instability.*
>
> *Like I said. Mum's the word!*

So error is immanent not only in the selection that determines the content and process of the assessment event, but also in the choice about what aspects will be elucidated in the assessment description.

Construction errors contain all of those errors in sampling, all the idiosyncrasies and biases that are contained in the construction of a specific test or set of demands that constitutes one element of the assessment event: these include not only the construction of the test content, of its elements, but also the construction of its form and style. Construction errors include all those generality errors attributable to the performance task itself, rather than to its timing or its context.

Practically, construction errors are indicated by all those differences in assessment description when the same construct is assessed independently by different people in different ways.

4. Labelling errors

Assessing is about describing some human performance. To give it a meaning the "some" must be specified: performance in typing; skill in mathematical problem solving; a dramatic presentation. So regardless of frame, it is necessary to specify in some way what it is that is being described. We must label the area of performance in some way, for otherwise it cannot be communicated.

The meaning of a communication is its reception, not its intention. In assessment the label is the message which is intended to describe a particular area of performance - involving particular knowledge, understandings, skills, processes, or whatever. The label has a particular meaning for the assessor growing out of this intention. Different meanings before the event will result in different assessment events being constructed to fit the label.

What meaning the assessed, or any other person who has access to the label, gives to it, is moot. But of one thing we may be certain. The meaning will not be identical to the meaning intended. The difference may be slight, or immense, but regardless of the magnitude will represent, at a fundamental level, an error (Korzybski 1933). Different meanings after the event will result in different interpretations of the assessment label, different inferences about what it implies.

An assessment must be an indicator of something. It must have a name. Differences in the meaning of the name, both before and after the event, constitute confusion and hence error. Labelling errors are defined by all the differences given to the meaning of the assessment (what it actually measures) by all the participants in the assessment event(s), and by the users of the assessment information.

Practically, labelling errors are indicated by the range of meanings given to the label by all those who use it before, during or after the assessment event.

5. Attachment errors

There is a further issue in regard to labelling. Once the label has been marked in some way, once the description is attached to it, where is it pinned? Does it belong to the person assessed? Is it more a description of the assessor? Does it represent some quantity or quality that might more appropriately hover somewhere in the space between, a relational field vector describing a complex interactional phenomena involving task, performance, assessor and assessed?

Given my ontological stance that all information is information about events, it follows that any attempt to attribute such information to a particular element of the event involves a fundamental epistemological error. To the extent that all other elements and conditions are held constant and overtly included in the description, to that extent is the simplification of language involved in the specific attribution partially justified; but such specificity of the conditions of the event tends at the same time to increase contextual error.

Attachment errors are the ontological slides that occur when a description of a relational event is attached to one of the elements of that event; specifically, when a complex relational event involving the construction of a test, an interaction of the test with a person, and a judgment of an assessor, is described as a property of the assessed person, this is an error in attachment.

Practically, attachment errors are indicated by the specification of those elements and boundaries of the assessment event that have become lost in the assessment description.

6. Frame of reference errors

Within the assessment arena are four competing definitions of the true, the correct, the impeccable. It follows that there are four associated notions of error. To the extent that the definitions of assessment truth, or more specifically the assumptions underlying them, are contradictory, so will be our methods for reducing error in the different frames; further, to the extent that the frames are confused, to that extent is error compounded. (See Chapter 13).

Frame of reference errors are defined by all those confusions and category differences that occur because of the different stable assumptions of the four frames of reference for assessment, as well as those contradictions and confusions that occur when shifts occur between frames during the assessment process.

Practically, frame of reference errors are indicated by specifying the frame in which the assessment is supposedly based, and indicating any slides or confusions that occur during the assessment events.

7. Instrument errors

Any measurement requires a measuring instrument. So any rank ordering, grading or scoring involves some measuring instrument; at the very least, such an instrument must attend to questions of calibration, which involves scale, replicability, and theory-practice bridging. Any claims to measurement must relate to some defined Standard scale. Whether the instrument is a test of some sort, or is assumed to have some material reality inside the mind of an examiner, all measuring instruments contain errors in mechanisms and hence in their readings. (See Chapter 9)

When psychometric theories are used, instrument errors are fed by all of the discrepancies between the theory and the empirical data, and are intrinsic in all of the notions of probability that pervade such theories.

Instrument errors then contain all those uncertainties of calibration, all those anomalies of replicability, all those confusions and discrepancies and mis-matches in theory-practice bridging, that are involved in the determination of the rank order, in the making of the mark, in the determination of the measure.

Practically, many aspects of instrument error are covered by other category errors. To avoid unnecessary overlap, I will limit the practical indicator of instrumental error to those errors implicit in the construction of the measuring instrument itself; what is conventionally called standard error of the estimate.

8. Categorisation errors

Any categorisation involves a comparison between a standard of acceptability, and a particular measurement or judgment about adequacy or quality.

Categorisation errors derive from confusions about the definition of standard of acceptability, from differences in the meaning of what is being assessed and in the magnitude of its measurement, and in the variability of the judgment process in which the comparison with the standard is made. (See Chapter 11)

Practically, categorisation errors are all those differences in assessment description that occur when particular data is compared with a particular standard to produce a categorisation of the assessed person.

9. Comparability errors

Comparability errors occur whenever assessment scores are added to produce a total score. Public examinations and grade point averages are examples of such summations, as are any qualitative assessments involving more than one criteria. What such additions mean, and who is privileged by such additions, are questions inherent in the process.

Comparability errors include all those confusions about meaning and privileging that inhabit the addition of test scores, grades or criteria related statements.

Practically, comparability errors are indicated by constructing different aggregates according to the competing models. The differences that these produce indicate the comparability error.

10. Prediction errors

Implicit in most assessment, and explicit in some, is the notion of prediction. Whilst the idea of generality contains some element of logic in its derivation, prediction can be pure magic - correlation without connection is very possible, and is not predicated on causal relationship. It has been reported that the number of storks sighted over London is correlated with the number of births in that city, and thus may be used as a predictor. The causal relation here is moot.

More seriously, many assessment descriptions are overtly or covertly connected to expectations about future performance. High school grades are presumed to be related to success at College or University. School performance is expected to relate to job success. Trade courses are designed to improve quality of performance in the workplace. So assessments on those courses might be expected to correlate with later performance. Yet even if they do, this in no way proves there is any causal link.

The criterion measures themselves are often problematic; most practical criterion measures themselves involve an assessment, subject to all of the sources of invalidity and error that dogged the original assessment. High predictive correlations may occur because both assessments are measuring something other that what they are described as measuring; for example, the ability to perform in competitive, written events, independent of the content. And low predictive correlations may mask genuine positive relationships because of all the errors entailed in the assessments, though such "genuine relationships" must forever be hidden, relegated to fantasy because divorced from empirical sustenance. Alternatively low correlations may mask the reality of relative homogeneity of performance status, or of genuine multi-dimensionality of that performance.

So interpreting the meaning of high correlations can be quite tricky. For example, if the rank order of students on a university entry examination in Physics correlates 0.9 with their first year Physics results at University this could be interpreted as an enormously successful outcome in terms of educational prediction. It is also completely consistent with the implication that no new Physics has been learnt, or that the University course has been completely unsuccessful in compensating for initial inequities in knowledge and opportunity.

What becomes apparent is that this area of prediction, which on the surface seems very amenable to empirical verification, is fraught with errors of interpretation which are neither measurable nor resolvable. Positioning and power relations will largely determine the trend of the discourse, and whether such discourse becomes a verification of validity, or an explication of error.

Explicit or implicit in most assessments is the claim that they relate to some future performance, that they predict a particular product from some future event, a quality of some future action. Prediction error is the extent to which these predictions, and the subsequent events, are not identical.

Practically, prediction error is indicated by the differences between what is predicted by the assessment data, and what is later assessed as the case in the predicted event.

11. Logical type errors

Test scores are often interpreted as giving specific information about what a student can or cannot do. For example, a score of 90 per cent on a spelling test gives no information about whether any individual item on the test was actually spelt correctly by a particular student. Any assumption to the contrary is a logical type error. Similarly, a score of 80 per cent on a mastery test gives no information about what information or skill has been mastered. Common inferences made from test scores are riddled with such logical type errors.

Logical type errors occur whenever there is confusion between statements about a class of events, and statements about individual items of that class.

Practically, logical type errors are made explicit when the explicit and implicit truth claims of a particular assessment are examined and any logical type errors are made explicit. Such exposure may invalidate such claims.

12. Value errors

All tests and examinations involve the construction of questions and the interpretation and valuation of answers. As such they are explicit and implicit statements about value; these particular questions, and these favoured answers, are implicit statements about what knowledge, actions, processes and interpretations are valued. And by implication, which are not so valued. Such implications move well beyond content; style and form and medium are of equal or more importance.

To the extent that the values implicit in the assessment event are not explicit, are contested, or are contradictory, to that extent is the assessment event invalid with respect to value. To the extent that the assessment event(s) and the event about which inferences are made are incongruent in terms of their value assumptions and emphases, to that extent is the error component engorged.

Practically, value errors are indicated by making explicit the value positions explicit or implicit in the various phases of the assessment event, including its consequences, and specifying any contradiction or confusion (difference) that is evident.

13. Consequential errors

Messick (1989a) and Cronbach (1988) both accept that the effects of testing have to be taken into account when assessing the validity of testing. It follows that any distortion of learning through the assessment process constitutes a source of error.

To take this view, however, is to make an extension to the meaning of validity, or of invalidity. For we have to ask, in what way does such distortion of learning detract from the appropriateness, usefulness, or meaningfulness of the inferences made from the test scores? Are the test scores less useful because they have distorted the learning process?

Certainly in such a situation the testing process has been counterproductive, which is a good reason for dismantling it, if learning is a major purpose of education. However, earlier chapters have shown this to be a naive proposition. Assessment has other more important if less salubrious social purposes.

Logically, distortion of learning increases error only if we take error to include not only the differences between what the test measures and what is or might be, but also between what the test measures and what might have been. This seems to take us into a rather transmogrified realm. Even so, any distortion of learning possibilities contributes to the violation of those persons whose learning, and possibility of growth, is thus diminished. And as that very learning is part of the event that the assessment presumes to measure, then it is legitimately included as inappropriate, and thus a source of error, a (retrospective) interactive interference effect.

Consequential errors involve all those negative effects on a student's learning and a teacher's teaching that are attributable to the assessment event. (To the extent that it produces inequity among sub-groups, positive effects on learning may also be involved).

Practically, at a simplistic level, consequential errors are indicated by the differential positive and negative effects that individual teachers and students attribute to the assessment process: At a more profound level it involves an explication of the very construction of their individuality, and all of the potentially violating consequences of those constructions. (See Chapters 4 & 5)

Invalidity according to Messick

Messick's (1989a) treatment of Validity in *Educational Measurement* is an excellent review of current (theoretical) state of the art, progressive in stance, and its implications vastly surpass current practice.

In this section Messick's work is looked at from the standpoint of invalidity, in order to indicate that the sources of invalidity indicated above are indeed well- established, if somewhat opaquely discerned, in the literature on validity.

Temporal error

Here are two passages from Messick that illustrate some of the temporal problems of validity. The first relates to the lack of necessary conjunction between construct meaning on the one hand, and stability of measure on the other:

> In regard to temporal generalizability, two aspects need to be distinguished: one for cross-sectional comparability of construct meaning across historical periods . . and the other for longitudinal continuity in construct meaning across age or developmental level. It should be noted that individual differences in test scores can correlate highly from one time to another (stability) whether the measure reflects the same construct on both occasions (continuity) or not. Similarly, scores can correlate negligibly from one time to another (instability), again regardless of whether the measure reflects the same or a different construct (discontinuity) on the two occasions (p57).

So even if the measure remains the same at different times, it may mean different things. And if the measure is different at different times, it may mean the same!

Here is the second example. Messick argues that it is not necessary to assume that

> the more generalizable a measure is, the more valid. This is not generally the case, however, as in the measurement of such constructs as mood, which fluctuates over time; or concrete operational thought, which typifies a typical developmental stage (p57).

From the standpoint of invalidity, that a test is invalid unless proved otherwise, how could the measurement of such an ephemeral quality ever be validated?

Contextual error

Contextual errors receive a lot of attention from Messick. Here is one example:

> Tests do not have reliabilities and validities, only test responses do. . . . test responses are a function not only of items, tasks, or stimulus conditions but of the persons responding and the context of measurement. This context includes factors in the environmental background as well as the assessment setting. . . . Thus,

> the extent to which a measure displays the same properties and patterns of relationships in different population groups and under different ecological conditions becomes a pervasive and perennial empirical question (p14-15).

This certainly captures the idea that the data belongs to a complex event, even though Messick does not follow through to the logical conclusion that the test score data cannot then be detached from the event and attached to an individual.

Moreover, in terms of error in individual measures he misses the point; for even with knowledge of the relationships between test measure - group - context, we still have no knowledge about the specific error in an individual score. (In group terms it could be anywhere within plus or minus three standard errors from the estimate).

Here is another example that raises the more fundamental issue of context as boundary condition:

> studies of the transportability of measures and findings from one context to another should focus on identifying all of the boundary variables that are a source of critical differences between the two contexts, as well as gauging the potency and direction of the effects of these boundary variables on events in the two conditions (p58).

Indeed, for science is nothing if it cannot adequately define the boundary conditions within which the limited experimental events that define its world can be controlled. So the assessment is invalid unless all the boundary conditions (that cause unexplained variance) can be specified. And, of course, they never can be.

Construction errors

Construction problems are often dealt with in terms of content validity. Messick comments that "the heart of the notion of so-called content validity is that the test items are samples of a behavioural domain or item universe about which inferences are to be drawn" (p36). He has some problems with this, for "to achieve representativeness . . . one must specify not only the domain boundaries but also the logical psychological subdivisions or facets of the behaviour or trait domain" (p39). Furthermore, "in point of fact, items are constructed, not sampled" (p40). And finally, Messick's crunch point :

> knowing that the test is an item sample from a circumscribed item universe merely tells us, at the most, that the test measures whatever the universe measures, and we have no evidence about what that might be, other than a rule for generating items of a particular type (p40).

So even the apparently simple task of getting some test questions together is fraught with difficulties, again justifying a invalidity label until compelling evidence is presented that these problems have been solved.

Labelling errors

Messick is adamant that "the meaning of the measure . . . must always be pursued - not only to support test interpretation but also to justify test use" (p17).

At least some of this meaning is carried by the construct label, and "constructs are broader conceptual categories than are test behaviours, and they carry with them into score interpretation . . . the evaluative overtones of the construct labels (p59).

One such problem with the label is how broad to make it. Messick spells out the dilemma:

> In choosing the appropriate breadth or level of generality for a construct and its label, one is buffeted by opposing counterpressures toward oversimplification on the one hand and overgeneralization on the other. . . . choices on this side (of oversimplification) sacrifice interpretative power and range of applicability as the construct might be defensibly viewed more broadly. At the other extreme is the apparent richness of high-level inferential labels such as intelligence, creativity, or introversion. Choices on this side suffer from the mischievous value consequences of untrammelled surplus meaning (p60).

Another problem with a label that applies to everybody is that different people do things in different ways:

> In numerous applications of these various techniques for studying process, it became clear that different individuals performed the same task in different ways

> and that even the same individual might perform in a different manner across items or on different occasions. . . that is, individuals differ consistently in their strategies and styles of task performance. . . this has consequences for the nature and sequence of processes involved in item responses and, hence, for the constructs implicated in test scores. . . test scores may mean different things for different people. . . for different individuals as a function of personal styles and intentions. . . Indeed, . . . a test's construct interpretation might need to vary from one type of person to another (p54-5).

So why not from one person to another? In this regard note that validity has always been a group concept. Human rights, with its associated absence of violence, is a term that applies to individuals and not to groups; to claim that 95 per cent of a population is not subjected to human rights violations such as torture, incarceration and extermination is hardly a claim for a good human rights record. Why is assessment any different?

It would seem from Messick's own example that the label must be individualised in meaning before it can validly be applied to an individual person.

Attachment errors

The idea that assessment data gives information about an event rather than about a person is contrary to the very conception of assessment in general, and to psychometrics in particular. However, there are glimmerings of light in Messick's work that are encouraging. Here are two examples:

> The possibility of context effects makes it clear that what is to validated is an interpretation of data arising from a specified procedure (p15).

> . . . the important validity principle embodied by this term (trait validity) might be mistakenly limited to the measurement of personal attributes when it applies as well to the measurement of object, situation and group characteristics (p15).

In the first quote the data is seen to be related to a procedure, that is, an event involving relationships; in the second the validity, if not the data, is seen clearly not to be limited to the personal.

Frame of reference errors

Messick does not mention frame of reference errors in the form that I have developed them in this dissertation. However he does talk of the various theoretical frameworks for intelligence, including the two well-known "geographic" models of intelligence as a single dimension, or as multiple discrete abilities. And then goes on to mention a computer model, an anthropological model, a sociological model and a political model. He then comments:

> If two intelligence theories sharing a common metaphorical perspective - such as uni-dimensional and multi-dimensional conceptions within the so-called geographical model - can engender the different world phenomenon of investigators talking past one another, as we have seen, just imagine the potential babble when more disparate models are juxtaposed (p61).

A close inspection of the literature on assessment obviates the necessity to imagine, for fact is indeed stranger than fiction, and indicates the massive sources of invalidity from this source.

Instrument errors

Instrument errors as such don't get much attention in this work, perhaps because, as I have defined them, they are an aspect of reliability rather than of validity, and so are dealt with in a different chapter in Educational measurement (Linn, 1989a).

However, he does note that "the very fact that one set of behaviours occurs in a test situation and the other outside the test situation introduces an instrument error"(p37), indicating that he is aware of a fundamental shift in context that pervades the use of tests for assessment.

Categorisation errors

About the validity of any particular categorisation Messick is remarkably silent. A short section on decision models of cost - benefits is all that scratches the surface of the chasm of silence (p78-80). This despite the fact that in practice the meaning of the categorisation assumes more importance than the meaning of the construct; to the individual student the distinction, or the failure, is more important than whether the assessment measured what it claimed to measure.

The substantiality of the standard is a necessary prerequisite to the allocation of a measure to a category: Or, for that matter, of the conversion of a category to a measure, as in a conversion of "better or worse" to "more or less." Are standards then irrelevant to construct validity, which in Messick's model is all validity? For surely the construct meaning given to a test score is submerged in the social world, in most cases, under the weight of its categorisation as a grade. To limit the definition of validity to test scores hardly affects the issue, because surely the categorisation then becomes the first interpretation, the first utility, the first action, and hence a crucial element in the validity discourse.

Should I really have been so surprised, as I most genuinely was when I realised for the first time, as I wrote the two preceding paragraphs, what had occurred? Was it a conscious decision on Messick's part not to include the categorisation issue in his extremely comprehensive study? Or is the erosion of the problem of the standard from professional and public memory so complete. Certain it is that though I have been very familiar with Messick's chapter for four years, and standards are my major area of interest, I had not noticed the almost complete omission of any treatment of the issue in his definitive paper on validity till now.

Whatever, categorisation errors remain a major source of invalidity in assessment, and without clear evidence to the contrary, must be assumed to be very large indeed, making most categorisations of individuals invalid.

Comparability errors

Now whilst Messick is certainly aware that "a single total score usually implies a unitary construct and vice versa" (p44), he does not develop many validity implications of this until he begins to discuss test-criterion relationships. He makes the point that "criterion measures must be evaluated like all measures in terms of their construct validity"(p70). He seems to accept that most criterion measures are "multiple and complex." He points out that it does not "make such sense logically to combine several relatively independent criterion measures . . . into a single composite as if they were all measuring different aspects of the same unitary phenomenon"(p74). He goes on to state that:

> On the contrary, the empirical multidimensionality of

> criterion measures indicates that success is not unitary for different persons on the same job or in the same educational program or, indeed, for the same person in different aspects of a job or program. furthermore, because two persons might achieve the same overall performance levels by different strategies or behavioural routes, it would seem logical to evaluate both treatments and individual differences in terms of multiple measures (p74-5).

Easy to say, of course, but much harder to do. Because this leads inevitably to the use of "judgmental weights that reflect the goals or values of the decision maker"(p75), which leads directly into all the confusions and errors dealt with in the chapter on comparability.

Prediction errors

Messick discusses prediction errors under the general rubric of test-criterion relations and decision making (p69 -88). He points out that "the major threats to criterion measurement . . . are basically the same as the threats to construct validity in general" (p73). In other words, errors are compounded in prediction errors because the errors in the test are multiplied by the errors in the criterion measure. In addition "other biasing factors include inequality of scale units on the criterion measure, which is a continual concern when ratings serve as criteria, and distortion due to improperly combining criterion elements into a composite" (p73). He talks of "inappropriate weights . . . applied to various elements in forming composites" (p73), yet who could say what an "appropriate" weight was?

So one source of confusion is whether the criterion domain "entails a single criterion or multiple criteria" (p74). He concludes that:

> use of measures of multiple criterion dimensions or components affords a workable approach to composite criterion prediction . . . by combining correlations between tests and separate criterion dimensions using judgmental weights that reflect the goals or values of the decision maker (p75).

Maybe, but this takes us into further sources of confusion related to differing values, differing goals, of different decision makers, and a concomitant further proliferation of error.

Value errors

In terms of the validity of tests, Messick is adamant that "the issue is no longer whether to take values into account, but how" (p58). It follows that "because validity and values go hand in hand, the value implications of score interpretation should be explicitly addressed as part of the validation process itself" (p59).

He is also clear that "data and values are intertwined in the concept of interpretation"(p16), and furthermore, "values . . . influence in more subtle and insidious ways the meanings and implications attributed to test scores with consequences not only for individuals but for institutions and society" (p59). So it is not only obvious biases expressed in interpretations that we are dealing with here, but "more subtle" mechanisms.

For example, not only are "some traits . . . open to conflicting value interpretations" (p60), (shouldn't this read "all traits"), but "the tenability of cause-effect implications is central, even if often tacitly, to the construct validation of a variety of educational and psychological measures such as those interpreted in terms of ability, intelligence, and motivation" (p58). So if cause-effect thinking is shown to be simplistic and epistemologically bankrupt in a more ecological world-view, where does that leave such "traits"?

So Messick centred his attention

> on the value implications of test names, construct labels, theories and ideologies, as well as on the need to take responsibility for these value implications in test interpretations. That is, the value implications, no less than the substantive or trait implications, of score-based inferences need to be supported empirically and justified rationally (p63).

Here Messick makes a brilliant case for the fundamental invalidity of all test data on the basis of value confusion and hence inability to interpret meaningfully test measures.

Consequential errors

Messick pays considerable attention to the consequential basis of test validity (p58-63). By this he means "the often subtle systematic effects of recurrent or regularised testing on institutional or societal functioning" (p18). He is firm that "social consequences cannot be ignored in considerations of validity" (p19). He then spells it out in more detail:

> The consequential basis of test interpretation is the appraisal of the value implications of the construct label, of the theory underlying test interpretation, and of the ideologies in which the theory is embedded. A central issue is whether or not the theoretical implications and the value implications of the test interpretation are commensurate (p20).

This may well be a central issue, but surely not the central issue. They may be commensurate and yet be utterly unequable to groups or to individuals. Messick himself acknowledges this later when discussing cost-benefit decision making:

> This concern with minimizing overpredictions, or the proportion of accepted individuals who prove unsatisfactory, is consistent with the traditional institutional values of efficiency in educational and personnel selection. But concern with minimizing underpredictions, or the proportion of rejected individuals who would succeed if given the opportunity is also an important social value in connection both with individual equity and with parity for minority and disadvantaged groups (p80).

Exactly, and Messick is equally precise when on the next page he concludes that "in practice, however, such balancing of needs and values comes down to a political resolution" (p81). That is, a solution based on power relations, which are inevitably asymmetrical. So if we are to be clear about invalidity errors of a consequential nature, we had best be mindful of the mechanisms through which such power relations are distributed and applied.

Messick's fudged solution

As briefly indicated above, Messick's chapter on Validity is a chamber of horrors, a gruelling journey through deep and varied sources of invalidity that would surely deter any rational person from ever attempting to show that any test was valid. Yet again and again he slides back into psychometrics, into "multiple choice" tests, into technological fixes, into the fudged solution.

Here is one such: "Tests," explains Messick, "are imperfect measures of constructs because they either leave out something that should be included according to the construct theory or else include something that should be left out, or both"(p34).

Not so. Messick has, conveniently, left out the fourth alternative, "or neither." And surely this is the alternative most congruent with his own analysis. By doing this he has assumed the very thing that is in doubt - that the construct can, in fact, be measured at all, in the light of epistemological issues, multi-dimensionality problems, value confusions, comparability errors, and so on.

Summary

To summarise, the notion of error is circumscribed by the construction of the event being described, just as it is boundaried by the epistemological assumptions of the judgment process.

Theoretically, error in assessment contains within its ambit all those ontological inadequacies, all those epistemological slides, all those logical contradictions, all those semantic obfuscations, all those definitional fudges, all those ideological camouflages, all those value variations, as well as all those potential empirical falsifications of implicit truth and accuracy claims, that characterise the field.

Practically, the description (measurement) of error is not dependent on any notion of a single truth, but rather on one of differences between multiple truths, all with some claim to legitimacy; these are implicit in the production of the assessment event, in the interpretations of the assessed and the assessor's experience of that event, including categorisations, and in the particular intended and received meaning of the communication of that judgment to others. The error becomes explicit when all of these phases of the assessment event are pluralised; when genuinely independent events are constructed; when independent categorisations are produced by participants in the event; when the judgments, and the meanings given to those judgments by involved persons, are compared.

Conclusion

Thus whilst the theoretical aspects of validity may indeed be fully discursive as Cherryholmes (1988) argues, the practical extent of invalidity is demonstrable as an empirical reality in the material world, partly as a result of that very discursiveness. For example the analysis presented earlier of the electrical automotive test presented irresolvable complexities in determining what empirical meaning could be given to the validity of the assessment. As the notion of validity is currently constructed, it would be resolved, if it was attended to at all, by the validity advocate giving an expert and coherent case for the defence, which would be unchallenged. That is, it would be resolved by resort to the Judge's frame of reference, and ignoring the other frames.

From the standpoint of invalidity, there is no such confusion. All of the suggested measures are useful measures, and the range of estimates that they produce for any one trainee indicates the range of error within which that person is being categorised. And we should not be surprised if at times this range covers the whole range of categories available.

As indicated in earlier chapters, the categorisation of persons has enormous effects on people, both in terms of their conceptions of themselves, and in their subsequent implicit and explicit exclusion from occupational opportunities. Such exclusion is not a discursive practice, but a very practical reality, though doubtless language is a significant factor in the acceptance of the violation. Further, the immense uncertainties associated with such categorisations is both demonstrable and measurable.

I have argued that validity discourse is currently constructed in such a way as to deny this demonstration. Invalidity discourse, based on the detailing of error components as presented here, is an advocacy for the defence of the examined rather than the examiner. As such it tends to redress the power imbalance, and hence reduce structural violence and increase social justice.

Return to Table of Contents

Part 6: Application

Chapter 18: Competencies, the great pretender
Chapter 19: National tests and university grades

Chapter 18: Competencies, the great pretender

Synopsis

In this Chapter, I apply the philosophical and conceptual positioning, tools of analysis, and the reconceptualised sources of error developed in this thesis to the competency based assessment policies and practices of Australia in the 1990s.

I first indicate how the notion of competency standards is overtly central to the whole competency movement, the introduction of which is shown to be overtly politically motivated. Thus the crucial links between political power and educational standards that are argued for in Chapters 3 and 4 become transparent.

I then go on to examine the validity, or more accurately, the invalidity of competency standards in the light of the thirteen sources of error specified in the previous chapter. The applicability of this analysis to a particular case is thus demonstrated.

Context: The re-birth of competencies in Australia

In the 1980s the discourse of politics became subsumed within the discourse of economics; quality of life was implicitly submerged, becoming a by-product of standard of living. And standard of living was explicitly defined by empirically derived statistics selected and interpreted by the theory of economic rationalism. Thus were the concomitant subjectivities of human misery, and the appalling atrocities of environmental degradation, excluded from the mainstream debate.

This same movement saw management practices move from control and exploitation of workers, modified at times by paternalistic concern for them, to a set of more manipulative practices described under the rubric of human resource management. This required the objectification of workers as a set of competencies, a necessary precursor to their ultimate replacement by more efficient computerised and robotic systems.

So the imposition of competencies as the basis of Australian technical and professional training during the late 1980s was in no way a decision informed by considered professional opinion; it was, from the start, an overtly political manoeuvre designed to solidify economic ideology in work practices, to demonstrate how skill and efficiency would reap their rewards in the "fair and just" game of the new internationally-competitive capitalist world order:

> The National Training Reform agenda is a co-operative national response to economic and industry restructuring, including labour marked imperatives and emerging requirements arising from workplace reform. The overriding aim is to increase the competitiveness and productivity of Australian industry, through industry responsive reform of the vocational education and training system. Flexibility to meet enterprise requirements within a stable and consistent national system is essential (National Training Board, 1992, p4).

The report goes on to state that "National competency standards provide the focal point of the new competency-based system" (p4). So here, quite explicit at the heart of the system, the manifest pivot, is the ubiquitous standard.

And of the essential arbitrariness of those standards or the necessary error in their measurement there is no word in this seventy one page report. Those two pillars of educational measurement, reliability and validity, do get a mention in the last page of the report. We are informed that assessment under the National Framework for the Recognition of Training (appropriately capitalised as a recognition of omnipotence) "provides for consistency as well as quality," and that one of the five principles of this approach is that "Assessment practices used shall be valid, ie. the techniques used must actually assess what they claim to assess." Furthermore, they must be reliable, in that "assessment approaches shall be able to be relied upon" (p71). And the Lord said, "Let it be done." And behold, it was done.

As Mc Donald (1994) nicely puts it: "The piece of commonsense that says that merely categorising something does not necessarily mean you can measure it easily, seems to have been lost" (p2).

In the first six pages of the NTB report, dealing with overview and context, the word "flexibility" appears eight times and "consistent" or "stable" six times. It is within this fundamental tension that the whole framework contradicts itself into nonsense. Let's unpack the argument in some detail:

> There is a clear need for a stable framework for national competency standards which is consistent across industries and across Australia. . . enabling nationally consistent assessment and certification to be achieved over time (p8).

Who needs this is unclear, but the rest is clear enough. The framework, which includes the ontological and epistemological assumptions about skills and learning and knowledge and the axiological assumptions about value, as well as the frame of reference about assessment, are all to be imposed on the basis of some "need." In other words, a centrally controlled system of education and training is to be imposed.

But there is to be flexibility. "Flexibility is required to enable specific industry and enterprise characteristics and necessary performance outcomes to be accommodated" (p8). The report goes on to indicate what is meant by flexibility; how flexibility is itself to be stabilised:

> This flexibility will be facilitated by the inclusion of general skills and knowledge in industry standards. Ensuring that industry standards look to the future, are packaged to allow multi-skilling, concentrate on important common skills and, where possible, not tied to particular forms of work organisation . . . simplicity as well as flexibility (p9).

In other words, flexibility is to be achieved by making the competence standards both general (non-specific) and simple.

So what have we got here? National Competency standards "provide the specification of the knowledge and skill and the application of that knowledge and skill to the standard of performance required in employment" (p9). And "assessment is the process of judging competency of an individual against prescribed standards of performance" (p11). So assessment is clearly in the Specific frame of reference, related to specific workplaces, related to particular jobs, related to performances that can be specifically described, and their levels clearly delineated and categorised (Norris, 1991).

On the other hand, the competencies are to be general, are to be non-specific, and furthermore have the truly remarkable quality of reflecting "not only industry's current but future needs"(p8), indicating perhaps a growth industry in astrology and clairvoyance training.

That the contradictions are so explicit gives hint to their genesis; they are the product of a succession of committees: Special Ministerial Conference on Training (1989), the Finn Report (1991), the Carmichael Report (1992), the Mayer Committee's Report (1992), and finally the input of the committees of the National Training Board.

At what level of discourse is all this? Are we involved in discourse at the rational-empirical level? Is this rhetoric really about measurement of competence? Or is this discourse at the mythical level? Is this about the construction of a national icon called competency standards around which a whole structure of power relations may be developed, and a whole new generation of workers constructed?

Porter (1992), speaking more specifically of the Carmichael report, sees it as

> a clever piece of policy writing, since its emphasis on diversity, options, pathways, and so on, obscures its desire to develop a training structure that is uniform, standardised and under the control of centralised bureaucracy (p54).

Similarly, Jackson (1992) is concerned that

> all of these reforms can be seen as a process of ideological capture, replacing the public purposes and social vision of education and other social institutions with the logic, and the social relations of, private wealth creation. The result is a profound and fundamental shift in where and how, and in whose interests, these institutions are controlled and managed(p159).

This is not to uncover a conspiracy to disenfranchise learners, teachers and small employers; Beevers (1993) explains that:

> In fact the Labor Party and the union movement in

> particular appear to have set out to do exactly the opposite. However the adoption of positivist, rationalist, bureaucratic and corporate managerial values and procedures has given rise to a curriculum model that - while providing advantages for politicians and systems managers - discriminates against the learning process and hence teachers, learners and small employers. . . What has been silenced is knowledge and skills that do not fit the technocratic, scientific, rationalist paradigm. The only knowledge and skills deemed worthwhile possessing are those believed to be directly related to increasing economic productivity (p103).

The paths to violence are indeed paved with good intentions.

Sources of error in competency assessment

In the remainder of this chapter I shall examine competency assessment in terms of the thirteen sources of invalidity conceptualised in Chapter 17.

Temporal errors

Firstly, there are the temporal errors in the criteria themselves. As Melton (1994) comments, "Inevitably the standards set reflect the perceptions of a particular group responding to perceived needs at a particular point in time. These will change as perceptions and needs change with time"(p288). Perhaps errors of this type are best categorised under construction errors, which become solidified in time because of the enormous structural and bureaucratic complexities involved in the production of the criteria of competency, which, despite what Melton says, do not usually specify a standard, a measurable level of adequacy.

Of more immediate concern are the variations in a particular person's performance over time. How are these to be interpreted? If competency is to be attached to the assessed person, then one adequate performance means the person has it, so is competent, so long as we assume that the thorny problem of adequacy has been solved. But if no adequate performance occurs, this could be a function of context, and does not necessarily imply incompetence. On the other hand, if competency is attached to events, then every performance ought to be adequate if the person is competent. Regardless of context? Confusion abounds!

In practice, temporal errors are confounded because no two events in which a human can engage can be identical, because time is change. Just as no two evaluations of a human event can be identical; they may involve identical categorisations, but the interpretative meaning of those categorisations change with time and with persons.

Potential confusions around the temporality of the measurement of competency standards abound; in current practice they are solved by pretending that they do not exist. They are a major source of error and confusion related to the meaning of any such measure.

Contextual errors

If performance depends on context, then assessment is about events in context, and competence is someone's judgment that a particular contextual behaviour is adequate. This is surely what "work" is all about, whether it is in school or on a personal project or in a paid job.

But this is so messy, because then a label can't be pinned on a person, because it belongs equally to a context. So how do we get back to a context-free categorisation? Easy:

> the assessment of competence is fundamentally about inferring competence from samples of performance. Under these circumstances, "competencies" are defined in terms of attributes, the competence is seen as deriving from the possession of, and ability to apply, relevant attributes to occupational tasks (Bowden, 1993, p55).

Of course, if fundamentally the assessment of competence is about inferring from samples of performance, then that's what you do, and the stuff about attributes is irrelevant. The attributes are politically necessary to get rid of the contextual error by assuming that there is none, so that the person can be categorised for all contexts.

Unfortunately, the empirical data contradicts the assumption, and makes the idea of such "attributes" very suspect, or at the least quite unmeasurable. Stanley (1993) sums up the current position:

> The message from the literature on transfer of training is that the idea of general strategies or competencies has been oversold. There are no substitutes for the building up of knowledge bases in specific domains. The evidence emerging from a number of recent cognitive studies is even stronger. It suggests that ways of thinking applicable for one domain of knowledge may be inapplicable in another (p145).

So the cost of attaching the categorisation to the person is to make it invalid in real contextual situations. The notion of competency standards solves the issue of context by fantasising the notion of an "attribute" called "competencies" which belong to the person so are independent of context. Reliability is thus increased in a psychometric sense. And validity is greatly diminished.

Construction errors

The original idea of competencies, in the Specific frame of reference, was to detail and teach all the little tasks that seemed to constitute the performance, and then test that they were all learnt to the required level of adequacy. The notion of competency standards as currently interpreted has moved a long way from this reductionist view. As Bowden (1993) explains it:

> the approach being taken to develop competency standards for the professions in Australia is not based on the professional's ability to perform specific tasks, but on the integration of relevant knowledge, skills and attitudes to complex workplace activities (p54).

Based, that is, on the measurement of knowledge of doubtful applicability and relevance, of skills that certainly have different applicability to different contexts, and of attitudes about which any inferences are surely problematic, and any measurement is highly suspect. So the price of solving the reductionist tiger has been to create an overgeneralised, undefined, unmeasurable and mis-attached elephant.

Melton (1991) elucidates the dilemma cogently:

> If competence is thought of as a deep structure of general ability then it is difficult to see how this abstract construct can be related to practice. It is close to offering a general theory of intelligence in forms of

> cognitive potential (p334).

It does indeed, and such a route is very rocky, as the last hundred years of controversy about intelligence tests have indicated.

Labelling errors

There are, as in all assessment systems, two types of labelling errors: There is the label of the particular competency; and there is the label of the categorisation of that competency.

As we move away from the Specific frame that can describe very specific eventful behaviours, we experience greater confusion in the meaning of the name that will become, in discourse, the referent for some practical competency that is ultimately defined either by some practical events in the world of work, or as some attribute or trait of a particular person. Regardless, whether we are talking of very generalised competencies such as "understands basic scientific principles," or very specific ones such as "adds two 2-digit numbers," what competency might mean in these domains is inevitably contested, and is different when viewed from different value positions or contexts, so the name will mean different things to different people. And this is not solved by the curriculum or test writer redefining such terms for their own purposes. As explained elsewhere, such a tactic may increase the reliability of the test, but it also increases its invalidity, because the user of the data generated from such tests is necessarily constrained to interpret the data in terms of the labels provided; labels to which they will attach their own meanings for their own purposes, and not magically absorb those constructed by remote curriculum and test constructors.

Similarly for the meaning of the label "adequacy" which is a necessary component of any discourse about competencies. Even if the problem of the meaning of the label that describes "what is being measured" could be solved, and we had a "scale" that was valid, we are still left with the problem of what is adequate along that scale; with the problem of the standard. This is also permeated with arbitrary and idiosyncratic definitions and interpretations, as well as enormous contextual variations; in short, another immense area of uncertainty, confusion, and hence error in personal categorisation and its interpretation.

Attachment errors

When competencies are described in terms of some particular assessor's evaluation of "adequate" work performance in a specific workplace, attachment errors are at a minimum - so long as competence is clearly tied to that particular work at that particular place by that particular person. Any reduction of the specification description, any attempt to attach the label to the person assessed, represents an attachment error, and, at least in the philosophical frame of this study, makes any competency claim ontologically invalid.

When such competence is reduced to a number of specific performances under specified conditions at specified levels of adequacy, attachment errors are at a minimum when all of this information is retained in the assessment description. Attempts to combine this information into one statement about competency, of which the specific behaviours are elements, is a logical type error which makes any competency claim logically invalid. Attempts to give a meaning to such a summation of elements involves both a comparability error, and an epistemological error in that the summation can have no meaning. Any such summation, by losing the contextual data related to the individual elements, results in an attachment error because the data now becomes attached to the person being assessed.

When, on the other hand, competence pretends to be some fixed attribute or skill or trait of the person examined, an attribute that is somehow "measured" by the person's interaction with a test, then the attachment error occurs when this measure is attached to other contexts, to other workplaces. It will then become apparent as contextual error or prediction error.

Frames of reference errors

Already the instability of the concepts of "competency," "competencies," and "competent" have been demonstrated. Norris (1991) comments that

> The requirement that competencies should be easy to understand, permit direct observation, be expressed as outcomes and be transferable from setting to setting suggests that they are straightforward, flexible and meet national as apposed to local standards . . . as tacit understandings of the words have been overtaken by the need to define precisely and operationalise concepts, the practical has become shrouded in

> theoretical confusion and the apparently simple has become profoundly complicated (p331).

He goes on to explicate:

> Behavioural constructs . . . express what is to be learnt in ways that make it transparent, observable and measurable. In contrast . . .the generic competency approach defines competence as broad clusters of abilities that are conceptually linked (p332).

In other words, the behavioural construct of competence is in the Specific frame, and the generic is in the General frame.

Messick (1984) confuses the issue further when he claims that competence is what a person knows and can do under ideal circumstances, whereas performance is what is actually done under existing circumstances. So competence is potential, is ability imminent. It follows that one successful performance demonstrates competence, because the conditions cannot be more than ideal, so one must assume they were less for any successful performance. On the other hand an unsuccessful performance can never demonstrate incompetence, because the conditions may not have been ideal.

So in theory there is confusion as to whether we are dealing with traits or demonstrated skills, concepts that require the General frame of reference, or particular defined behaviours, which require the Specific frame. In practice the confusion proliferates, for invariably the description of the standards that define the cut-off is either non-existent or vague, as indeed is the measuring instrument or instruments which will provide data to which the standard must be compared. So the practical assessment of what is adequate must be made in the Responsive frame - an intuitive response from the assessor. Such "subjective" admissions are, of course, utterly inadmissible, for the success of the whole charade is dependent on the appearance of objective accuracy and precision. Luckily, this is possible if the assessment mode shifts to the Judge's frame. So this is what happens, and certainty is reestablished, albeit in a different frame than that theoretically intended.

To summarise, analysis of competency assessment in terms of frames of reference indicates semantic chaos, discourse riddled with self contradictions. Out of it all there still emerges, from all involved, belief that the system works. And in as much as people are categorised, it does indeed work. To further believe however that some accurate measure of minute error has emerged from such conceptual confusion and personal lack of awareness is to substitute blind faith for rational thought.

Invalidity from this source is thus profound, and stems from the epistemological irrationality that must occur when frames of reference with contradictory assumptions are amalgamated without distinction into a single discourse.

Where does all this leave the individual student? Apparently presented with a list of clearly defined outcomes, things to know and do at predetermined levels of competence, closer inspection leaves the student with a list of ambiguous topic headings and ill-defined "skills," on the basis of which he or she will be tested, and then categorised by comparison with opaque standards visible only to the professional eye of the teacher. Was it ever otherwise?

Instrumental errors

Referring to standardised and/or criterion referenced tests, Berlak (1992) notes that "The credibility of these tests depends upon the claim that they are scientific instruments" (p181). Just so the credibility of competency assessment as a whole. The notion that these assessment systems are based on the measurement of clearly defined standards is what provides the educational, moral and public relations glue that transforms a set of fragile value and assumption struts into a powerful cognitive structure.

Yet it is surely a false claim. There are rarely such standards available, even at the practical level. At the level of physical factory products, standards that are related to some criteria of quality can sometimes be set up and measured, but these are a far cry from the "attributes" that predate competence in personal performance.

As described in the section on frame of reference errors, the whole discourse is emersed in epistemological confusion. What is important to note here, however, is that by pretending to belong to the Specific frame, the professional necessity to provide estimates of standard errors of measurement, necessary in the General frame, is side-stepped; not that educational practice ever paid much attention to that professional necessity.

The instrument, as apposed to any test, thus is firmly inside the mind of the assessor, an intuitive judgment hidden beneath the overt scientism of the competency label with its overtones of specific behaviours and definable standards.

Categorisation errors

Competencies must be described and then categorised. To categorise a competency we must first measure it and then compare the measure to a standard. As a result of this comparison we may then categorise the performance as adequate or inadequate, or the person as having, or not having, the competence.

So can we measure accurately these competencies that are described? Norris (1991) comments :

> there is a massive mismatch between the appealing language of precision that surrounds competency of performance-based programmes and the imprecise, approximate and often arbitrary character of testing when applied to human capabilities (p337).

As to the standards, these are normally presented as criteria to consider, as hints to decision makers, rather than defining the point on the measure that dichotomises a continuity. And even if there was a scale or measurement, and so the "standard" could be specified, how could it ever be anything other than arbitrary? A political decision based on data permeated with individual subjectivity and value.

Levin (1978) described the use of minimal outcomes in schools in the United States. It applies equally to the use of competencies in Australia in the 1990s:

> we do not have the knowledge bases to construct a defensible set of performance standards for certifying student competencies except in the most arbitrary sense. Whether such arbitrary standards are worthwhile in themselves may be debatable. Their inability to predict with any confidence that which is important in adult life is not debatable (p314).

Comparability errors

Melton (1994) accurately describes the sort of processes that are actually involved in competency assessment:

> Assessment is not simply a matter of ticking off

> whether individuals can or cannot perform tasks to certain clearly defined levels. Rather it is about looking at evidence, and making judgments about the levels of competence achieved based on the evidence provided. The evidence may be gathered from a variety of sources including observation of performance in the place of work, observation of specially set tasks, records of tasks that the candidate has performed in the past and from questioning the candidate on any aspect of the performance. Clearly much judgment needs to be brought to bear in interpreting such a range of evidence (p288).

And, of course, a judge will give a particular weight to a particular source of evidence, and will give a particular interpretation to the data available from each source, so that the meaning of any such final judgment must be quite obtuse, and different from the meaning given by another judge, even if the categorisation is the same, which seems unlikely in most cases.

Prediction errors

Because competencies in Australia have been specifically politically invoked to improve work practices and hence profitability in industry, prediction errors occur when the produced competencies do not specifically do all of those things; it is possible, remotely so in my view, that the educational events wrapped around competency standards might indeed in some cases have some validity in regard to the first of these claims, related to work practices, though some early research does not support this (Gillis, 1995). Of course, even if there is some correlation between the competence measure and some later predicted outcome, this in itself does not indicate causal link between the two categorisations that is mediated through the competency attribute.

In fact, as I have argued in the section on Consequential errors, it is unlikely that any empirical data will be collected in this regard because it is the assumption on which the whole scheme is premised, and thus not amenable to investigation.

Logical type errors

In all of its cyclic incarnations, competencies as specific behaviours have invariably encountered the criticism that they are reductionist, that they fragment knowledge, that they are in essence, trivial. Perhaps it is sufficient here to give two references:

> It cannot be assumed that mastery of the elements of competence will automatically lead to the achievement of more complex skills in the higher reaches of the hierarchy (Melton, 1994, p188).
>
> If I were to place competence within the art of pottery which I practise. Seeing it wholistically from the perspective of a great tradition of planetary and historical scope, I would only say: competence, your name is mud. (Beittel, 1984, p119).

In the Australian context, competencies face an identity crisis in that they are uncertain whether they are to be interpreted as holistic summations of such specific behavioural elements, or as specific behavioural outcomes of holistic mental attributes.

If the former, then the logical type error occurs in the summation, in the confusion of members of a class (the specific behaviours) with the whole class (the competency). In the latter case the logical type error occurs in the confusion of a description of a class (the generic competency) with members of that class (specific context-related work performances).

Either way confusion is confounded and error escalated through the attempt to define and describe competencies in any place other than their area of actual performance.

Value errors

Competence implies some purposeful act; a person is competent when she does something adequately in some context. So the first question to be asked in a competency judgment is: What ought the person do in order to be deemed adequate? This is not a factual question, but a value premise. And it is where every list of competencies must begin. Pearson (1984) argues that "until the value premise is made the competency claim cannot get off the ground" (p34). Thus all competency descriptions are based on value premises, which are usually unstated.

One implication of this is, as Norris (1991) points out, that "In the effort to describe competence in precise, transparent and observable terms, to predict the specific outcome of effective action, what is in fact happening is the pre-determining of good practice" (p334).

To the extent that competency requirements dictate school programmes, they also determine that "The measure of success that is applied for the schools is not the degree to which they foster intrinsically meaningful activities, but the degree to which they satisfy competence-related outcomes (Levin, 1978, p311). Levin (1978) goes on to assert that "Certification standards are signals to the schools of what is considered important by society, and their message will not be lost in individual teacher decisions or organizational ones" (p314).

Jackson (1993) perceives that the underlying intent of competency based teaching and assessment is to provide more governmental control on teaching institutions, and any effect on individual learning is secondary to this:

> the achievement of competency-based curriculum may not be about lasting improvement in individual performance at all, but about making teaching and testing accountable to a standard through a warrantable set of procedures. Technically, it is not the competence of the individual which is assured by these methods, but the competence of instruction and the liability of the institution. The shift is central to the power and sophistication of the competency paradigm as a tool of governance and an ideological force (p157).

How are these values transmitted to the individual student? What is the value learning that accrues? Here is a world of learning presented with machine-like crispness, sets of facts and relations and skills as neat as a computer board; the world of learning and of work reduced to packaged modules to be eaten up and deposited in the appropriate mental filing cabinet for later reproduction at so many dollars an hour.

Yet as we have seen, this whole operation begins from a particular view, generally not stated explicitly, of best practice: a particular positioning; a particular attachment to certain sorts of power and affect relations; a particular consciousness about work and its effects; and a begging of the question of who benefits from these particulars.

Where does the individual student position himself in this value matrix? He is supposedly acquiring the competencies that will allow flexibility in various job performances. Yet his experience may deny the usefulness and relevance of what is being presented. Even so, the competencies must be achieved. So rather than flexibility, such a student will learn not flexibility, but conformity; not a producer of new work practices, but a consumer of old ones.

Invalidity in terms of value then derives not only from the bias that derives from unstated value assumptions, but from the very specificity of stated intentions, and their contradiction by associated social effects; that is, by those very contradictions that are at the heart of symbolic violence.

Consequential errors

Elsewhere in this dissertation I have argued the centrality of assessment procedures to the construction of the individual in society. Commenting on the scene in the United States, Berlak (1992) comments:

> Among all assessment procedures, standardised and criterion-referenced tests are particularly privileged, that is, they serve as the single most powerful regulators of schooling practice, shaping the language used in public discussions about schooling, the criteria for judging the competence of students, and the range of possibilities considered for reforming the schools (p194).

And Jackson (1993) sees Australia in the 1990s following along a similar path:

> the discourse of competency increasingly defines not only our current practice but also the parameters of our imagination on issues of education and training (p159).

So here is one clear consequence of the competency movement. Increasingly the boundaries of discourse become narrower, and the possibilities for diversity become constrained, as notions of specifiable behaviours, performances, outcomes, skills and abilities, all defined by persons outside the training institutions, begin to dominate educational discourse. There is the further mythical belief that in some magical way standards are incorporated into these competency descriptions, which can be precisely measured and compared to such standards.

Students in this context are cogs in a gigantic machine. They are disempowered in terms of the substance and the value assumptions that predate what is to be learnt. There is no notion here of learning that grows out of specific purposes, learning styles or values of students, or of curricula negotiated to meet such purposes. Nor indeed is there any sense of relatively autonomous teaching agencies offering, among them, a proliferation of solutions to the relatively intractable problems of job training. As presented, competency assessment is the solution. The problems, whatever they may be, have been pre-empted. The job of training is to implement the solution. The function of evaluation is to indicate that the solution has been implemented. The closed black and white fantasy circle is complete.

Invalidity in terms of consequences stems most profoundly from the loss of the initial problem, which has been firmly removed outside the closed circle of competency discourse. For the National Training Board (1992), "the overriding aim is to increase the competitiveness and productivity of Australian industry"(p4), an aim now subsumed under the solution, which is assumed to be the National competency assessment system; so within the discourse of competency assessment not only can this question about productivity not be answered, it cannot even be asked, because its answer is itself.

For the individual student the potential errors in categorisation are immense. This particularly applies to students already enmeshed in work practises. Their learning cannot be based on local analyses of work environment deficiencies, or on creative transformations of work practices, because it is dedicated to their absorption of pre-ordained competencies which are supposed to magically provide such solutions. And if (when) the magic doesn't work, there is no place to go, for success has too long been dependent on the acceptance of absurdity.

Summary

I have argued that there are at least thirteen sources of invalidity that affect the measurement of competency standards. I contend that any one of these, applied to the assessment of individual students, would make the assessment of that student in these terms invalid.

Return to Table of Contents

Chapter 19: National tests and university grades

Synopsis

In this chapter I apply the reconceptualised notion of invalidity to national literacy testing, and to the definitions of grades within my own university.

These are presented as specific examples of the potency of the invalidity conceptualisation.

National Literacy Testing

Context

In its edition of 15-16 March, 1997, the newspaper Weekend Australian announced on the front page under the heading "All pupils face tests of literacy" that:

> The literacy and numeracy of every Year 3 and 5 student will be tested from next year under a historic agreement between the Commonwealth, States and Territories yesterday.
>
> The Catholic and independent schools sectors have indicated they will support the national testing program, which will be linked to uniform education standards to measure the reading, writing, spelling and mathematical ability of students.
>
> . . . The federal Minister for Schools, Vocational Education and Training, Dr Kemp, described the literacy strategy as a "historic agreement for the children of Australia" because it stresses that every child starting school from next year will be able to read, spell, and add up within four years.

The literacy test is to be based on that developed some years ago by the NSW Education Department, and it is this test to which the following critique is addressed, in terms of the thirteen sources of invalidity.

Temporal errors

Temporal errors are indicated by the differences in assessment description when the assessment occurs at different times.

No estimates of temporal errors in the national literacy testing program exist. They would, of course be easy to obtain and would be small compared to some of the other sources mentioned here. Small, that is, for most students. But the same theory that predicts this also predicts that a small percentage of students (randomly placed and unfindable) would have large discrepancies. But even small discrepancies would destroy the notion of infallibility that seems to be necessary for such tests to be publicly acceptable. This is what test administrators call public confidence, and I have more accurately named a psychometric fudge.

Contextual errors

Contextual errors include all those differences in performance and its assessment that occur when the context of the assessment event changes.

Literacy is a concept of great educational importance, of diffuse and contested and multi-dimensional meaning. It involves at the very least reading and writing. Yet reading what under what conditions? And writing what under what conditions? A test defines the what and defines the conditions: Tightly specifying the conditions improves the reliability; yet at the same time it obviously disguises and increases the lack of generality and hence increases the contextual invalidity.

Essentially, the context of test-taking is not the context in which literacy, in most of its forms, is demonstrated.

Construction errors

Construction errors are indicated by all those differences in assessment description when the same construct is assessed independently by different people in different ways, whilst the broader context of the assessment is held constant.

It would be relatively simple to take samples of children and have teachers and researchers and the children themselves make independent assessments of various aspects of their literacy, and estimate construction errors by comparing the estimates with each other and with the result of the test. This writer has no doubt that such an experiment would presage the immediate cessation of such testing.

Labelling errors

An assessment must be an indicator of something. It must have a name. Differences in the meaning of the name, both before and after the event, constitute confusion and hence error. Labelling errors are defined by all the differences given to the meaning of the assessment (what it actually measures) by all the participants in the assessment event(s), and by the users of the assessment information.

Literacy tests presume to measure literacy. But which particular aspects? What could any test score tell us about any of those aspects? What meaning is given to those aspects by any particular teacher? How does that meaning compare to that teacher's concept of literacy? And what action could be taken by any such teacher on the basis of those meanings to help any child more than that teacher is currently helping? The extent to which these questions produce diffuse and varied and contradictory answers gives an indication of labelling error. And the meaning of literacy includes such confusion. The problem is not solved by imposing a definition; this enables us to increase reliability, and reduce the apparent error in measurement. But it is a reductionist trick, a semantic scam. The concept of literacy is diffuse, so any attempt to measure to is, at best, extremely imprecise, and, at worst, meaningless and hence impossible.

Attachment errors

Attachment errors are the ontological slides that occur when a description of a relational event is attached to one of the elements of that event; specifically, when a complex relational event involving the construction of a test, an interaction of the test with a person, and a judgment of an assessor, is described as a property of the assessed person, this is an error in attachment.

The implications of this source of invalidity for literacy testing are immense. Any information about the test cannot be unattached from the particular test and attached to the student as a "trait" or "ability." This involves a demystification of the whole process and its highly suspect theoretical underpinning. Such demystification relates it to the fundamental question "What do we really know about where this literacy score came from?" The answer is clear. A particular group of people selected a particular set of multiple choice test items which the student answered under particular conditions and were subsequently given a score which placed them in a rank order and some of them were then classified as below a standard which did not exist until this group or another group were so classified.

The point to emphasise here is that the score does not belong to the student. It belongs to the experimental event of which the student was a part. Any movement beyond this point requires another experiment - which, of course, produces another event, with concomitant multiplication of confusion and error.

Frame of reference errors

Practically, frame of reference errors are indicated by specifying the frame in which the assessment is supposedly based, indicating the errors according to its own and other frames, and indicating any slides or confusions that occur during the assessment events.

In testing programs on literacy the tests pretend to be in the Specific frame of reference. The tests are talked about as though there are clearly defined and accepted specific tasks which students must do successfully in order to be considered literate or numerate. And that there is some predefined standard to which appeal may be made. Neither of these claims are true. The test items which are the basis of complex statistical manipulations are subjectively chosen by test constructors from the pool available, which may include some that they themselves specifically construct. And there is no standard other than that defined by the test itself. Some test constructors talk of an absolute scale. They are deluding themselves (Behar, 1983). All test data are based on item statistics which are norm referenced from groups of test takers. So the tests produce a rank order of merit and the test controllers (test makers, educational administrators, or political funders), make arbitrary decisions about adequacy (Glass, 1978). What we can be certain of is that the tests will produce a rank order in which some students will obtain higher scores than others. That is what the tests are designed to produce. Any implications beyond this about adequacy are arbitrary value judgments.

Instrument errors

Instrumental error is implicit in the construction of the measuring instrument itself; what is conventionally called standard error of the estimate, or is indicated by the spread of judgments of independent assessors about a particular performance on a particular test.

One assumes that in national literacy tests this relatively small source of error (simple reliability) will be known to test constructors, forgotten by test administrators, and withheld from teachers and students. Regardless, such an estimate of error gives no information about the error of a particular student, and withhold the statistical information that only two thirds of actual students will have "true" scores within these limits, and as the total numbers tested increase, an increased number of individual students will be given completely unacceptable estimates.

At a more fundamental level, the instrument (the test) cannot measure anything because there is no Standard, no adequate theory -practice bridging to define the scale, no scale, and thus no measure that the scale may proscribe, that may subsequently be compared to a standard of acceptability.

Categorisation errors

Categorisation errors derive from confusions about the definition of standard of acceptability, from differences in the meaning of what is being assessed and in the magnitude of its measurement, and in the variability of the judgment process in which the comparison with the standard is made.

Practically, categorisation errors are all those differences in assessment description that occur when particular data is compared with a particular standard to produce a categorisation of the assessed person.

The implications of this for literacy testing are profound. For not only is the meaning of the score highly suspect, but there is in fact no standard of literacy with which such a score may be compared. The standard is an arbitrary point selected after the event by the test makers and is based on the particular test, or on the particular items used in the construction of the test. Such circularity in definition produces a closed system that is the stuff of fantasy, but not of scientific measurement.

Comparability errors

Comparability errors include all those confusions about meaning and privileging that inhabit the addition of test items, test scores or grades. Practically, comparability errors are indicated by constructing different summaries or summations according to competing models. The differences that these produce indicate the comparability error.

Literacy is a multi-dimensional concept. As such, a single dimensional scale can be used to measure the concept, but such a measure could not be given a meaning. In particular, any categorisation (involving a standard, assuming one exists) cannot be given a meaning, because it could never be certified whether any particular single - dimensional score was above or below that "standard." Because such meaning is central to the notion of validity, such inability to give a meaning makes any uni- dimensional test of literacy constitutionally invalid.

Prediction errors

Practically, prediction error is indicated by the differences between what is predicted (or more subtly implied) by the assessment data, and what is later assessed as the case in the predicted event.

There is an implication in the national literacy program that the scores show that some children are illiterate, and that without special intervention triggered by this test they will remain illiterate. Such an implication could be empirically tested, assuming there was some satisfactory definition of illiterate. I know of no such definition, or of any program to develop one or otherwise empirically test the effects of the testing.

Logical type errors

Logical type errors occur whenever there is confusion between statements about a class of events, and statements about individual items of that class. Practically, logical type errors are made explicit when the explicit and implicit truth claims of a particular assessment are examined and any logical type errors are made explicit. Such exposure may invalidate such claims.

In a rare burst of intellectual honesty the earlier versions of the literacy test were headed "Aspects of literacy" (NSW, 1995). Such a test cannot be a test of literacy. Statements about some members of the class do not apply to the whole class. All literacy and numeracy tests have this problem. They are essentially a summation of the specific items that the test comprises, and assumptions cannot be made of implications beyond this. Psychometrics could be defined as a statistical sampling game that produces a fantasy about traits in order to sidestep the contradictions that flow from the reality that all test scores are summations of discrete elements, and that all information about the individual elements is lost in the summation.

Value errors

Value errors are indicated by making explicit the value positions explicit or implicit in the various phases of the assessment event, including its consequences, and specifying any contradiction or confusion (difference) that is evident.

The National tests purport to give information about individual students that might lead to remedial action. The value appealed to is that of helping students and improving performance. The tests are not diagnostic and so give no information about what particular misconceptions or problems (if any) particular students may have, apart from the extremely error-prone response from one or two items. Even if such diagnosis were available, its usefulness would depend on teachers being able to use it to improve student performance. And since it is not known whether or not teachers have already targeted some children for extra attention, its usefulness would depend on whether the test produces the same group for special attention, and in cases of difference whether the National test produced a more valid selection.

As there is no evidence that the tests will help children, it may be less naive to suggest that the main value behind the test is to help politicians gain prestige by appearing to solve a problem (which may not exist).

Consequential errors

Consequential errors are indicated by the differential positive and negative effects that individual teachers and students attribute to the assessment process. At a more profound level the test may involve an explication of the very construction of their individuality, and all of the potentially violating consequences of such constructions.

The focus of the testing will be on those who are lower in the rank order. Theoretically these will be identified, and will improve as a result of special instruction. The magical improvement kit has not yet been produced, so such consequences are doubtful, especially as literacy (as most people understand the term), is so dependent on a whole range of experiences outside the school. What is more certain as a consequence is that such students will be classified as "failures" or "remedial" and will, in many cases, construct their individuality accordingly.

Summary

Practically, the description (measurement) of a person's literacy is not dependent on any notion of a single truth, but rather on one of differences between multiple truths, all with some claim to legitimacy; these are implicit in the production of the assessment event, in the interpretations of the assessed and the assessor's experience of that event, including categorisations, and in the particular intended and received meaning of the communication of that judgment to others. The error becomes explicit when all of these phases of the assessment event are specified; when genuinely independent events are constructed; when independent categorisations are produced by participants in the event; when the judgments, and the meanings given to those judgments by involved persons, are compared.

When such errors, contradictions, and confusions are acknowledged, the pristine purity of the test score disappears, to be replaced by a wide fuzzy band of possibilities; then rank orders recede, standards evaporate, categorisations are exposed as fantasy, and the whole inane and monstrous structure crumbles.

National literacy tests have thirteen charges (at least) to answer before being considered valid. Many of these are so fundamental that I doubt any reputable educator would take the case.

University grades

Context

Just as honesty begins with self, so truthfulness should not ignore the home campus. My own university has announced a new grading system for the categorisation of students (Flinders University of South Australia, 1997). An analysis of the grade descriptions indicates six criteria are used. A summary of the descriptors is given in Table 1 (see next page).

In the next section I will examine this grading system in terms of the thirteen sources of invalidity.

Temporal errors

If grades refer to a particular race that students have competed in, then temporal errors need not concern us. Description of the event includes a particular time and place and tomorrow is another day. If, on the other hand, they are presumed to indicate some skill or competency of the student, then they must also be presumed to have some constancy over time. Tomorrow is the same day in terms of traits and capacities and skills and understandings. At an ideological level the whole exercise depends on this. So if "skills" are developing then logically only the most recent performance should count. And if they are not developing then what are the students learning?

Table 1 Grade descriptions

Grade	core work	knowledge, competency	texts	wider reading	debates approaches	original and creative
pass 2 50-54	undertaken	adequate	basic		some familiarity	
pass 1 55-64	more	sound	sound		good general level of familiarity	
credit 65-74	additional	sound	sound	done	apply a range	
distinction 75-84	considerable additional	advanced	advanced	considerable	broad familiarity and facility at applying	developing a capacity
high distinction 85-100	considerable additional	highest level	in depth	extensive	highest level of proficiency in applying	combining knowledge with
fail 0-49	fail to complete	fail to demonstrate				

Further to this, if they have actually learnt through the process of doing the project, or through any subsequent feedback, then the product becomes invalid because the state of the student is now different from that state when the product was produced, and another temporal error has been perpetrated.

In this sense, tests are premised on an assumption of student stasis; the more the student learns during or subsequent to any test information, the more that test information becomes outdated and hence in error.

Contextual errors

The grade descriptors do not mention context. But they imply a range of possible contexts, assessment modes, media, and processes. In order to make sense of such grades we must infer that the performances on which they are based are independent of the context in which they are produced; that is, they must represent a fixed measurable property of the student rather than a particular response to contextual events. It has been argued in this thesis that to believe this is an ontological error. Regardless, it is obvious that human behaviour, including cognitive behaviour, varies markedly according to context, so to reduce contextual error of the grades it would be necessary to specify the context of the events resulting in students' products, and the events resulting in the assessors product (the grade).

Without such contextual specification therefore the grades must of necessity be invalid.

Labelling errors

There are two labels; the label that describes the measure, and the label that describes what is measured. The assumption of these descriptors is that the measure can exist independent of what is measured. That grades have a reality independent of what is being graded. That administrative convenience can become a substantive reality. As indeed it will. But at what cost to professional integrity or student justice?

And even if this assumption is not nonsense, there is still the problem of the meaning of the grade. As I have indicated, the grade demarcations are so vague that errors within each criteria must be immense. Further, once the criteria become combined into a single dimension all information about individual criteria is lost, so all meaning related to the criteria likewise dissolves.

Attachment errors

As I have reiterated in many places in many ways in this thesis, information gained from tests is information about an event in which an individual student is an element. Any attempt to attach the description or data to the student, rather than to the total event, is an ontological slide. Attempts to not only attach to the student, but to some particular conceptual entity which the student is fantasised to have, takes us even deeper into the ontological bog. Error is reduced as the completeness of the event is recaptured. Such recapturing, of course, nullifies the use of simple numerical and graded categories.

In this case we have, in terms of the definitions of the grades, at least six independent classification events, all of which are supposed to contribute to the final grade. Error is indicated by any differences or confusions of grade within or among such events.

Frame of reference errors

The criteria would appear to indicate the Specific frame. Within each criteria there are indicators of grade demarcations. However, these are hardly adequate for specifying any standards. What is the difference between basic, sound, advanced, and in-depth? How do you draw fine lines between some familiarity, good general level of familiarity, broad familiarity and facility at applying? And how do you differentiate between developing a capacity for creativity, and combining knowledge with creativity? How else would you know a capacity was being developed than by relating it to knowledge? Obviously within the specific frame the indicators for cut-offs are hopelessly inadequate, and in this frame the system is grossly invalid.

Perhaps though this is unfair. Perhaps it is only political fashion that has forced this appearance of competency. The word "highest" appears twice, and this is obviously a normative term belonging to the General frame. Yet there are no percentiles given for grade boundaries, so standards are not possible to define within this frame. There are of course marks given that are appropriate to each grade. The Calender makes it clear, or at least implies strongly, that these marks are awarded as subdivisions of the grade, rather than that the grades are based on some previously determined marks. What is done in practice is moot. Regardless, the system is unworkable in the General frame, because there are no guidelines in this frame to decide grade boundaries. Within this frame therefore immense errors of miscategorisation must be expected.

In the Judge's frame, where as the reader will recall there is no error by definition, there is no problem. There never is. Judges have no problem differentiating between more core work, additional core work, and considerable additional core work. Even when, as appears to be the general case from the descriptions of courses given in the Calender, no core is specified. Or even, indeed, between the different "soundness" that differentiates pass level 1 from credit when applied to sound knowledge and competencies, and the sound understanding of texts.

It seems apparent that the criteria here are a competency smoke screen, a vague set of hints that allow assessors to continue to do what they have traditionally done; create a comparative order of merit of doubtful meaning , and at the same time allocate rather arbitrary grade boundaries to the rank order. The specification of criteria, naive and inadequate as they are, nevertheless fortifies the "scientism" of the Judge's frame, armouring its uncertain certainties with a coating of current assessment dogma.

Instrumental errors

With a plethora of assessment modes-assignments, practical work, observations, tests, examinations, it is sometimes difficult to actually locate the instrument, the "objective" machine that makes the measure. And of course there is no such objective machine. The fantasy that tests of various kinds are measuring instruments unfortunately remains a prevailing myth in the assessment of persons. The assessment modes are merely techniques used to fix a performance in time and space, to give it reality through some semblance of permanency. This allows, at least theoretically, independent judgments to be made of their "quality" or relative merit.

In practice the actual instrument, the place where the standard resides, the conceptual theory-practise link is established, the mark is produced, the comparisons are made, and the categorisations established -- all of these exist inside the mind of the examiner. So there is no objective instrument, and the assessment is clearly in the responsive mode, subject to all the normal variations and anomalies of idiosyncratic subjective judgments. Single examiners, which is the norm for university assessments, disguises this reality by nullifying in advance all competing judgments.

Categorisation errors

Within each criteria the categorisation boundaries are defined by words or phrases of extraordinary vagueness and imprecision, when it is remembered that this purports to be the official description of the categories that determine students' futures.

For example, assuming that the "core work" for a particular course has been precisely defined, then it might indeed be possible to determine whether it had been "undertaken." Or even if "more" than the required work was done, meriting a pass 1 classification. But how to distinguish this "more" from the "additional" core work required for a credit, or the "considerable additional" work required for a distinction or a high distinction, is unspecified. And how does the "sound" knowledge and competency required for a pass 1 differ from the "sound" knowledge and competency required for a credit, and in what way is that different from the "advanced" knowledge and competency required for a distinction or the "highest level" required for a high distinction? Surely it would be easier to be honest and say: "Rank order the students somehow and then draw arbitrary grade boundaries!"

Comparability errors

How are estimates for different criteria to be summated? The meaning of the final grade can only have a meaning in relation to the criteria if the loadings for each criteria are transparent, for how can we compare grades if they can mean different things. And how can we compare them anyway? How does "developing a capacity for original and creative work" in Commercial Law B compare with the same description in Human Resource Management or Mathematics 1A or Cognitive Science? What could "developing a capacity" possibly mean in any context, for that matter? And how can you compare the core work between subjects when it isn't specified in most cases? Indeed, if it isn't specified in some detail the whole grade description structure is entirely unworkable within a subject, for how could "additional" be judged without knowing what it was additional to?

Prediction errors

Whilst there are no overt predictions made in terms of these grades, there are some covert ones of immense significance. Certainly entry to higher degree programs is largely determined by the grades obtained, so there is an implicit prediction that students with lower grades are less suited to such further work. And, of course, students who fail are predicted as unsuited to qualify for work in particular fields.

Performance in academic course work, even if it could be accurately assessed, is very different from performance in professional work contexts. Yet the former is often, and increasingly, a necessary prerequisite for the latter. So the predictive validity of the grades would seem to be of vital importance, especially in those professions that require academic qualifications.

As indicated in Chapter 15, predictions about job performance on the basis of any selection criteria tend to be very low indeed, and correlations of 0.3 are considered very adequate. That this is ten percent better than pure chance indicates the immensity of the predictive error, and the extraordinary extent of the social injustice perpetrated through such mechanisms.

Logical type errors

Referral to Table 1 indicates there are six elements to the class of each grade. Are all elements required for the grade to be awarded? Or are five out of six enough? Or is one element enough for a higher grade? Could a person graded pass 2 be at a high distinction in five elements and be categorised pass because they had not done wider reading? How would we know that? If the elements must all be attained for a given level of grade then necessarily the lowest level in any element will alone determine the grade. If individual common sense gives the answers to these questions what can grades mean when common sense is so disparate?

Attention to possible logical type errors of this sort indicate inevitable massive confusion and thus error in the interpretation of these grades.

Value errors

What are some of the value errors implicit in this system? An obvious one is that "more and less" is synonymous with "better and worse." This shows very clearly in the descriptors for core work, knowledge and competency, and wider reading. The clear implication of these columns is that more is better.

This has considerable social as well as semantic significance. There is a value clearly implied that students should do more work than is specified or required, and that merit is accumulated through such activity. There are uncomfortable parallels here with current work practices in a competitive market, where workers are increasingly expected to work longer hours for no additional remuneration, and this exploitation becomes twisted by ideology to become a symptom of professionalism.

Another value, whose implications influence comparability errors, is that of terribly ordered learning. The six criteria must march along in unison otherwise they are unusable. It seems, for example, that original and creative thinking can only occur after masses of core, and additional conceptual work, has been understood. Is this true? Cannot innovative practical methodologies be constructed with very little specific knowledge? Cannot original and creative practical experiments and equipment design be produced to specifications with almost no knowledge of background theory? The limiting of the terms original and creative to the top two grades involves very prejudicial assumptions.

Consequential errors

How quickly and how intensely do students accept the judgments of their assessors as to the relative merit and idiosyncratic opinion (disguised as absolute value) of their academic performance? To what extent is the camouflage of error, the appearance of certainty, a predominant factor in this acceptance? To what extent does such acceptance affect later work, either positively or negatively? To what extent is the academic student constructed by the apparently objective measurements of their grades?

Such effects may be consistent within discernible sub-groups of students, or may be individually differentiated. Regardless, the questions indicate a particular category of invalidity, and in fairness to all students demand answers if the extent of invalidity for this criteria is to be explicated.

Not a problem

Does the confusion with its attendant error that is evident here create a problem for assessment in academia? It would seem not. Hopeless as the descriptors are, they are probably no better or worse than those they replaced, nor of others elsewhere. Academics just do not seem to problematise confusion and error in the measurement of "standards," at least not in academic discourse.

Is validity an issue? I checked the journal Assessment and Evaluation in Higher Education. Of a total of 195 articles in this journal from 1986 to 1996 only nine dealt, directly or by implication, with the problem of error, or inconsistency, or lack of validity in grading or marking. Of these nine there were three articles on validity which did not deal with inconsistency or error as any sort of a problem or issue. Four dealt with marker reliability, and two of these trivialised the notion of error in their conclusions.

Closer to home, Orrel's (1997) examination of the thinking-in-assessment of "everyday academics" revealed sometimes some angst in assigning a grade, but little concern that the "standard" itself might be illusory. And she commented that "A notable silence in the academic's discourse was any reference to the considerable technical measures that exist for assuring validity and reliability in assessment"(p397). But then, as they were clearly in the Judges frame of reference, such comment would have constituted a mind-shattering contradiction.

Conclusion

In the vernacular, it's a matter of "no worries, mate, business as usual!"

Return to Table of Contents

Part 7: Concluding statement

Chapter 20: Out of the fog

Chapter 20: Out of the fog

This study was begun to answer one fundamental question: How is error in measurement of standards obscured in most practical events involving assessment of persons?

Before I commenced work on this thesis I had already worked on this particular aspect for two years, and had written about ten chapters for a book on the subject. Further work during the past two years at Flinders University has developed and enlarged the scope of the work. As well, I have traversed some side roads, taken some wrong turnings, and come to a few dead ends. For example, at one stage it seemed the whole focus of the work would be on competencies. At another point interviews with assessment experts, administrators, teachers and students loomed large on the agenda. So at various times I was diverted from the main topic but always returned to it, often with fresh insights.

Tying the focus to the concepts of validity and invalidity was a relatively late development, only possible after the literature on validity was reframed as an advocacy for the test taker. The centrality of comparability to the whole assessment issue was similarly a late discovery.

I am personally pleased at the outcome. I can now make some sense out of what seemed non-sense; I have shown how some of the fudging was accomplished, and why it was important, in terms of social stability, to do so. At the same time I have, I believe, forged a powerful tool for the analysis of invalidity of assessment, and hence of error in the categorisation of individual persons--a tool based on a shift in positioning from test giver to test taker.

In a rational world the thirteen sources of invalidity, developed in many cases by reframing and repositioning the accepted scholarship in the field of assessment, should be sufficient to halt the conceptual blindness, the blatant suppression of error, the subtle fudges, and the myth of certainty that permeates the "science" and expertise of categorising people. Full acceptance and individual specification of even one of these sources could revolutionise current practice. However, as the study indicates, the world in which assessment resides is far from that rational world to which much of the writing in this thesis appeals.

I have tried to be clear about some of the forces that work on all of us that will encourage the reader to react strongly and negatively to many of my arguments, to dismiss them as anathema. The work is immoral in that it conceptually threatens the inviolability of standards and their measurement, a lynch pin of the cultural production of the modern individual. And it is revolutionary in that action based on its conclusions would destabilise to a point of destruction many, probably most, educational and work practices that result in the categorisation of people.

On the other hand, the basic contentions of this project are not contentious at the top levels of evaluation in Education, Medicine, or Law: Ph D theses in Education are assessed by different examiners and it is expected that such assessors will often differ in their judgments of quality; when expert opinions are sought in medicine both diagnosis and treatment prescriptions may differ markedly; and the seven judges in the high court often give conflicting verdicts.

The work could be criticised as being unduly negative. Even if the claims of the thesis are true, or partially true, is its position not destructively unhelpful? We need to categorise people, so take away the standard and what remains? How can people live with the certainty of uncertainty? At the very least, give us an alternative. And whilst I have not developed the alternatives, I have certainly presented them. The Responsive frame has many developed modes of assessment within its boundaries. The chapter on quality clearly indicates one way to go. We live in a world of complexity and uncertainty, a fuzzy multi-dimensional world of immense variety and diverse interpretations. What is challenged in this work is the myth that this complexity can be reduced to simple linear dimension by some sort of examination, as a preliminary to comparing with some standard of adequacy somewhere defined.

This thesis does not contend that people cannot be pinpointed along such dimensions, butterflies permanently fixed on the board. It happens to millions every day. What is shown is that such categorisations are inevitably permeated with confusion, uncertainty and error, that genuine rather than fudged estimates of much of this error can be made, and that this particular violation of the human mind and spirit will continue until they are.

Return to Table of Contents

References

(APA) American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington: American Psychological Association.

Apple, M. (1982). *Education and power*. Boston: Routledge and Kegan Paul.

Arendt, H. (1969). *On violence*. London: Penguin.

Australian National Training Authority. (1994). *Towards a skilled Australia: A national strategy for vocational education and training* : Australian National Training Authority.

Ayers, W. (1993). *To teach: The journey of a teacher*. New York: Teachers College Press.

Ball, S. (1994). *Education reform*. Buckingham: Open University Press.

Barone, T. (1992). Beyond theory and method: A case of critical storytelling. *Theory into Practice*, 31(2), 143-146.

Barton, L., Whitty, G., Miles, S., & Forlong, J. (1994). Teacher education and teacher professionalism in England: some emerging issues. *British Journal of Sociology in education, 15(4)*, 529-543.

Bateson, G. (1972). *Steps to an ecology of mind*. New York: Ballantine Books.

Bateson, G. (1979). *Mind and nature*. London: Wildwood House.

Becker, H. (1990). Generalising from case studies. In E. Eisner & A. Peshkin (Eds.), *Qualitative enquiry in education: the continuing debate*. New York: Teachers College, Columbia University.

Beevers, B. (1993). Competency-based training in TAFE: Rhetoric and reality. In C. Collins (Ed.), *Competencies: the competencies debate in Australian education and training*, . Canberra: Australian College of Education.

Behar, I. (1983). *Achievement Testing*. Beverly Hills: Sage Publications.

Beittel, K. (1984). Great swamp fires I have known: Competence and the hermeneutics of qualitative experiencing. In E. Short (Ed.), *Competence: Inquiries into its meaning and acquisition in educational settings*, (pp. 105-122). Lanham, MD: University Press of America.

Benett, Y. (1993). The validity and reliability of assessments and self-assessments of work-based learning. *Assessment and Evaluation in Higher Education, 18(2)*, 83-93.

Berk, R. (1986). A consumers guide to setting performance standards on criterion reference tests. *Review of Educational Research, 56(1)*, 137-172.

Biesta, G. (1994). Education as practical intersubjectivity: towards a critical-pragmatic understanding of education. *Educational Theory, 44(3)*, 299-317.

Bloom,B. (Ed.). (1956). *Taxonomy of educational objectives; Handbook 1, cognitive domain*. New York: David Mc Kay.

Bloom, B. (1976). *Human characteristics and school learning*. New York: McGraw-Hill.

Bloom, B., Hastings, J., & Madaus, G. (1964). *Handbook on formative and summative evaluation of student learning*. New York: McGraw Hill.

Borthwick, A. (1993). Key competencies - Uncovering the bridge between the general and vocational. In C. Collins (Ed.), *Competencies: the competencies debate in Australian education and training*, . Canberra: Australian College of Education.

Bourdieu, P., & Passeron, J. (1977). *Reproduction in education, society and culture*. London: SAGE Publications.

Bowden, J., & Masters, G. (1993). *Implications for higher education of a competency-based approach to education and training*. Canberra: Australian Government Publishing Service.

Bracht, G., & Glass, G. (1968). The external validity of experiments. *American Educational Research Journal, 5(4)*, 437-474.

Broadfoot, P. (Ed.). (1984). *Selection, certification and control*. London: The Falmer Press.

Brown, R. (1973). *Religion and violence*. Philadelphia: The Westminster Press.

Bruner, J. (1986). *Actual minds, possible worlds*. Cambridge: Harvard University Press.

Bucke, R. (1969). *Cosmic consciousness*. New York: E.P. Dutton Co.

Burchell, G., Gordon, C., & Miller, P. (Eds.). (1991). *The Foucault effect*. London: Harvester Wheatsheaf.

Burgess, R. (Ed.). (1985). *Issues in educational research: Qualitative methods*. London: The Falmer Press.

Burton, N. (1978). Societal standards. *Journal of Educational Measurement, 15(4)*, 263-273.

Cairns, L. (1992). Competency-based education: Nostradamus's nostrum. *The Journal of Teaching Practice, 12(1)*, 1-32.

Camera, H. (1971). *Spiral of violence*. London: Sheed and Ward.

Campbell, J. (1956). *Hero with a thousand faces*. New York: Meridian Books.

Carr, W., & Kemmis, S. (1983). *Becoming critical*. Geelong: Deakin University Press.

Cherryholmes, C. (1988). *Power and criticism*. New York: Teachers College Press.

Cherryholmes, C. H. (1988). Construct validity and the discourses of research. *American Journal of Education, 96(3)*, 421-457.

Clough, E. E., Davis, P., & Sumner, R. (1984). *Assessing pupils: a study of policy and practice*. Windsor: NFER-Nelson.

Codd, J. (1985). *Curriculum discourse: text and context*. Paper presented at the National Conference of the Australian Curriculum Studies Association, La Trobe University, Melbourne.

Codd, J. (1988). The construction and deconstruction of educational policy documents. *Journal of Education Policy, 3(3)*, 235-247.

Collins, C. (Ed.). (1993). *Competencies: the competencies debate in Australian education and training*. Canberra: Australian College of Education.

Collins, R. (1979). *The credential society*. Orlando: Academic Press Inc.

Cox, R., & 1965. (1965). *Examinations and higher education: A survey of the literature*. London: Society for Research into Higher Education.

Cresswell, M. (1995). Technical and educational implications of using public examinations for selection to higher education. In T. Kellaghan (Ed.), *Admission to higher education*, . Dublin: Educational Research Centre.

Cronbach, L. (1969,). *Validation of educational measures*. Paper presented at the The 1969 invitational conference on testing problems:

Towards a theory of achievement measurement.

Cronbach, L. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity*, (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L., Rajaratman, N., & Gleser, G. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology, XVI(2).*

Cronbach, L. J. (1990). *Essentials of psychological testing*. (Fifth ed.). New York: Harper and Row.

Delandshere, G., & Petrosky, A. (1994). Capturing teachers' knowledge: performance assessment. *Educational Researcher, 23(5),* 11-18.

Docking, R. (1995, January 1995). Competency: What it means and how you know it has been achieved. *NTB Network- Special Conference Edition, 18.*

Donmeyer, R. (1990). Generalizability and the general case study. In E. Eisner & A. Peshkin (Eds.), *Qualitative enquiry in education*, . New York: Teachers College, Colombia University.

Downs, C. (1995). *Key competencies: A useful agent for change?* . Richmond: National Centre for Competency Based Assessment and Training.

Eisner, E. (1988). The primacy of experience and the politics of method. *Educational Researcher, 17(3),* 15-20.

Eisner, E. (1990). The meaning of alternative paradigms. In E. Guba (Ed.), *The paradigm dialog*, . Newbury Park: Sage Publications.

Eisner, E. (1991b). Taking a second look: Educational connoisseurship revisited. In M. McLaughlin & D. Phillips (Eds.), *Evaluation and education at quarter century*, (pp. 169-187). Chigago: The National Society for the Study of Education.

Eisner, E., & Peshkin, A. (Eds.). (1990). *Qualitative enquiry in education.*

Eisner, E. W. (1985). *The educational imagination*. (second ed.). New York: Macmillan.

Eisner, E. W. (1991). *The enlightened eye*. New York: Macmillan.

Fay, B. (1987). *Critical social science*. New York: Cornel University Press.

Feyerabend, P. (1988). *Against method*. London: Verso.

Finn, B. C. (1991). *Young people's participation in post-compulsory education and training* . Canberra: Australian Educational Council Review Committee.

Fish, S. (1980). *Is there a text in the class? The authority of interpretive communities*. Cambridge, Ma.: Harvard University Press.

Foucault, M. (1972). *The archaeology of knowledge*. London: Tavistock Publications.

Foucault, M. (1982a). Questions of method: an interview with Michel Foucault. *Ideology and Consciousness, 8(6)*, 3-14.

Foucault, M. (1982b). The subject and the power. In H. Dreyfus & P. Rabinow (Eds.), *Michel Foucault: Beyond structuralism and hermeneutics*, . Brighton: Harvester.

Foucault, M. (1988). *Politics, philosophy, culture: interviews and other writing*. New York: Routledge.

Foucault, M. (1992). *Discipline and punish*. London: Penguin.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18(9)*, 27-32.

Freud, S. (1963). *Civilisation and its discontents*. London: The Hogarth Press.

Friedenberg, E. (1969). , *Proceedings of the 1969 invitational conference on testing problems*, . Princeton: Educational Testing Service.

Garcia, G. E., & Pearson, P. D. (1994). Assessment and diversity, *Review of research in education*, (Vol. 20, pp. 337-391).

Garman, N. (1994). Qualitative enquiry: meaning and menace for educational researchers. In J. Smyth (Ed.), *Qualitative approaches in educational research*, (pp. 3-14). Adelaide: Flinders University of South Australia.

Garman, N., & Holland, P. (1995). the rhetoric of school reform reports: sacred, sceptical and cynical interpretations. In R. Ginsberg & D. Plank (Eds.), *Commissions, reports, reforms and educational policy*. Westport: Praeger.

Gillis, S., & Macpherson, C. (1995,). *Examination of the links between pre-employment qualifications and on the job competency based assessment.* Paper presented at the Australian Association for Research in Education, 25th Annual Conference, Hobart.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18,* 519-521.

Glass, G. (1978). Standards and criteria. *Journal of Educational Measurement, 15(4)*, 237-261.

Golstein, H. (1979). Changing educational standards: A fruitless search. *Journal of the NAIEA,* 11(3), 18-19.

Gonzalez, E. J., & Beaton, A. E. (1994). The determination of cut scores for standards. In A. C. Tuijnman & T. N. Postlethwaite (Eds.), *Monitoring the standards of education.*

Good, F., & M, C. (1988). Grade awarding judgements in differential examinations. *British Educational Research Journal, 14(3)*, 263-281.

Green, M. (1994). Epistemology and Educational Research: the Influence of Recent approaches to Knowledge. *Review of Research in Education, 20*, 423-464.

Green, P. (1981). *The pursuit of inequality*. Oxford: Martin Robertson.

Guba, E. (1990). *The paradigm dialog*. Newbury Park: SAGE Publications.

Guilford, J. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*, 427-439.

Hacking, I. (1991). How should we do the history of statistics? In G. Burchell, C. Gordon, & P. Miller (Eds.), *The Foucault effect*, . London: Harvester Wheatsheaf.

Haertel E H. (1991). New forms of teacher assessment, *Review of Research in Education*, (Vol. 17, pp. 3-29).

Hambleton, R., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Nijhoff Publishing.

Hambleton, R., & Zaal, J. (Eds.). (1991). *Advances in educational and psychological testing*. Boston: Kluwer Academic Publishers.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement, Third edition*, . New York: American Council on Education, Macmillan Publishing Company.

Hartog, P., & Rhodes, E. (1936). *The marks of examiners*. London: Macmillan and Co.

Harvey, L., & Greed, D. (1993). Defining quality. *Assessment and*

Evaluation in Higher Education, 18(1), 9-34.

Horkheimer, M., & Adorno, T. (1972). *Dialectic of enlightenment*. New York: Herder and Herder.

House, E. (1991). Evaluation and social justice. In M. McLaughlin & D. Phillips (Eds.), *Evaluation and education at quarter century*, (pp. 233-247). Chicago: University of Chicago Press.

Howe, K. R. (1994). Standards, assessment, and equality of educational opportunity. *Educational Researcher, 23(8)*, 27-33.

Hulin, C., Drasgow, F., & Parsons, C. (1983). *Item response theory: Application to psychological measurement*. Homewood, Illinios: Dow Jones-Irwin.

Huxley, A. (1950). *The perennial philosophy*. London: Chatto and Windus.

Illich, I. (1971). *Deschooling society*: Calder and Boyers Ltd.

Jackson, N. (1993). Competence: A game of smoke and mirrors? In C. Collins (Ed.), *Competencies: The competencies debate in Australian education and training*, . Canberra: Australian College of Education.

Jaeger, R., & Tittle, C. (Eds.). (1980). *Minimum competency achievement testing*. Berkeley: McCutchen Publishing Corporation.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement, Third edition*. New York: American Council on Education, Macmillan Publishing Company.

Johnston, B., & Dowdy, S. (1988). *Teaching and assessing in a negotiated curriculum*. Melbourne: Robert Anderson and Ass.

Johnston, B., & Pope, A. (1988). *Principles and practice of student assessment*. Adelaide: South Australian Education Department.

Jones, L. (1971). The nature of measurement. In R. Thorndike (Ed.), *Educational measurement: second edition*, (pp. 335-355). Washington: American Council on Education.

Kavan, R. (1985). *Love and freedom*. London: Grafton Books.

Keeney, B. (1983). *Aesthetics of change*. New York: The Guilford Press.

Kennedy, K., Marland, P., & Sturman, A. (1995). *Implementing national curriculum statements and profiles: corporate federalism in retreat*. Paper presented at the Annual Conference of the Australian Association for Research in Education, Hobart, 26-30 November.

Knight, B. (1992). Theoretical and practical approaches to evaluating the reliability and dependability of national curriculum test outcomes, : Unpublished article.

Korzybski, A. (1933). *Science and sanity*. Lakeville, Con: International non-Aristotelian Pub. Co.

Laing, R. (1967). *The politics of experience*. Harmondsworth: Penguin.

Lather, P. (1991). *Getting Smart: Feminist research and pedagogy with/in the postmodern*. New York: Routledge.

Lazarus, M. (1981). *Goodbye to excellence: A critical look at minimum competency testing*. Boulder: Westview Press.

LeCompte, M., Millroy, W., & Preissle, J. (1992). *The handbook of qualitative research in education*. San Diego: Academic Press Inc.

Levin, H. (1978). Educational performance standards: Image or substance. *Journal of Educational Measurement, 15(4)*, 309-319.

Lincoln, Y. (1990). The making of a constructivist. In E. Guba (Ed.), *The paradigm dialog*. Newbury Park: Sage Publications.

Lincoln, Y. (1995). Emerging criteria for quality in qualitative and interpretative research. *Qualitative Inquiry, 1*(3275-289).

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23(9),* 4-14.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher, 20(8),* 15-21.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Lorge, I. (1951). The fundamental nature of measurement. In E. Lindquist (Ed.), *Educational measurement,* (pp. 533-559). Washington: American Council on Education.

Madaus, G. F. (1986). Measurement specialists: Testing the faith - A reply to Mehrens. *Educational Measurement: Issues and Practice, 5(4),* 11-14.

Mager, R. (1962). *Preparing instructional objectives.* Palo Alto, CA: Feardon Publishers.

Marshall, C. (1990). Goodness criteria. In E. Guba (Ed.), *The paradigm dialog*, . Newbury Park: SAGE Publications.

Masson, J. (1991). *Final analysis*. London: Harper Collins.

Masters, G. (1994, 17 March). *Setting and measuring performance standards for student achievement.* Paper presented at the Public Investment in School Education: Costs and Outcomes, Canberra.

Maturana, H., & Guiloff, G. (1980). The quest for the intelligence of intelligence. *Journal of Social Biological Structures, 3.*

Mayer, C. C. (1992). *Putting general education to work: the key competencies report* . Melbourne: Australian Educational Council and Ministers of Vocational Education, Employment and Training.

McDonald, R. (1994, October, 1994). Led astray by competence? Paper presented at the Australian National Training Authority, Brisbane.

McGovern, K. (1992). National competency standards - the role of the National Office of Overseas Skills Recognition. *The Journal of Teaching Practice, 12(1)*, 33-46.

Meadmore, D. (1993). The production of individuality through examination. *British Journal of Sociology in Education, 14(1)*, 59-73.

Meadmore, D. (1995). Linking goals of governmentality with policies of assessment. *Assessment in Education, 2(1)*, 9-22.

Melton, R. (1994). Competencies in perspective. *Educational Research, 36(3)*, 285-294.

Messick, S. (1989a). Validity. In R. L. Linn (Ed.), *Educational Measurement, Third edition*, . New York: American Council on Education, Macmillan Publishing Company.

Messick, S. (1989b). Meaning and values in test validation. *Educational Researcher, 18(2)*, 5-11.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23(2)*, 13-23.

Miller, A. (1983). *For your own good*. New York: Farrar, Straus, Giroux.

Miller, A. (1984). *Thou shalt not be aware*. London: Pluto Press.

Miller, C., & Parlett, M. (1974). *Up to the mark: a study of the* *examination game*. London: Society for Research into Higher Education.

Millman, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational Measurement, Third edition*, . New York: American Council on Education, Macmillan Publishing company.

Mishler, E. (1986). *Research interviewing*. Cambridge: Harvard University Press.

Mitroff, I., & Sagasti, F. (1973). Epistemology as general systems theory: An approach to the design of complex decision-making experiments. *Philosophy of the social sciences(3)*, 117-134.

Moss, P. A. (1992). Shifting concepts of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62(3)*, 229-258.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23(2),* 5-12.

Mykhalovskiy, E. (1996). Reconsidering Table Talk: Critical thoughts on the relationship between sociology, autobiography and self-indulgence. *Qualitative Sociology, 19(1),* 131-151.

Nairn, A. (1980). *The reign of ETS* . Washington.

National Training Board. (1992). *National Competency Standards: Policy and Guidelines (Second Edition)* . Canberra: National Training board.

National Training Board. (1995, January 1995). Who's doing what? Assessment in Australia today. *NTB Network - Special Conference Edition,* 19-20.

Norris, N. (1991). The trouble with competence. *Cambridge Journal of Education, 21(3),* 331-341.

Nuttall, D. (1979). The myth of comparability. *Journal of the NAIEA, 11(3),* 16-18.

Nuttall, D., Backhouse, J., & Willmott, A. (1974). *Comparability of standards between subjects.* (Vol. 29). Oxford: Evans/Methuen Educational.

Oakley, A. (1991). Interviewing women. In H. Roberts (Ed.), *Doing feminist research*, . London: Routledge and Kegan Paul.

Orrell, J. (1996). Assessment in higher education: an examination of everyday academic's thinking-in-assessment, beliefs-about-assessment, and a comparison of assessment behaviours and beliefs. Unpublished Ph D, Flinders University of South Australia, Adelaide.

Partington, J. (1994). Double-marking students' work. *Assessment and Evaluation in Higher Education, 19(1),* 57-60.

Pawson, R. (1989). *A measure of measures.* London: Routledge.

Pearson, A. (1984). Competence: a normative analysis. In E. Short (Ed.), *Competence; Inquiries into its meaning and acquisition in educational settings,* (pp. 31-40). Lanham, MD: University Press of America.

Pennycuick, D., & Murphy, R. (1988). *The impact of graded tests.* London: The Falmer Press.

Perkins, D., & Salomon, G. (1988). Teaching for transfer. *Educational Leadership*(September), 22-32.

Persig, R. (1975). *Zen and the art of motorcycle maintenance: An enquiry into values.* New York: Bantam Press.

Persig, R. (1991). *Lila: An enquiry into morals.* London: Bantam Press.

Peters, M. (1996). *Poststructuralism, politics and education.* Westport: Bergin & Garvey.

Phillips, D. (1990). Subjectivity and objectivity: an objective enquiry. In E. Eisner & A. Peshkin (Eds.), *Qualitative enquiry in education.* New York: Teachers college, Colombia University.

Popkewitz, T. (1984). *Paradigm and ideology in educational research.* London: The Falmer Press.

Porter, P., Rizvi, F., Knight, J., & Lingard, R. (1992). Competencies for a clever country: Building a house of cards? *Unicorn, 18(3),* 50-58.

Prigogine, I., & Stengers, I. (1985). *Order out of chaos.* London: Fontana.

Quine, W. (1953). *From a logical point of view.* New York: Harper and Row.

Rechter, B., & Wilson, N. (1968). Examining for university entrance in Australia: Current practices. *Quarterly Review of Australian Education, 2(2).*

Reilly, R., & Chao, G. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 33(1),* 1-55.

Resnick, D. P., & Resnick, L. B. (1985). Standards, curriculum and performance: A historical and comparative perspective. *Educational Researcher, 14(4)*, 5-20.

Rorty, R. (1991). *Objectivity, relativism, and truth.* Cambridge: Cambridge University Press.

Rose, N. (1990). *Governing the soul: The shaping of the private self.* London: Routledge.

Rosenberg. (1967). *On quality in art: criteria of excellence, past and present*. Princeton: Princeton University Press.

Royal Commission. (1974). *Report on the suspension of a high school student* . Adelaide: South Australian Government.

Sadler, D. R. (1987). Specifying and Promulgating Achievement Standards. *Oxford Review of Education, 13(2)*, 191-209.

Sadler, R. (1995). Comparability of assessments, grades and qualifications. Paper presented at the AARE Conference, Hobart, 24 November.

Schmidt, F., Hunter, J., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: a red herring. *Journal of Applied Psychology, 66(2)*, 166-185.

Schnell, J. (1980). *The fate of the earth*. London: Picador.

Schwandt, T. (1990). Paths to enquiry in the social disciplines. In E. Guba (Ed.), *The paradigm dialog*, . Newbury Park: SAGE Publications.

Scriven, M. (1991). *Evaluation thesaurus; fourth edition*. Newbury Park, Cal: SAGE Publications.

Shepard, L. (1991). Psychometricians' beliefs about learning. *Educational Researcher, 20(7)*, 2-16.

Shepard, L. A. (1993). Evaluating test validity, *Review of research in education, 19*.

Sherman, R., & Webb, R. (1988). *Qualitative research in education*. New York: Falmer.

Slater, P. (1966). *Microcosm*. New York: John Wiley.

Smith, B. (1994). Addressing the delusion of relevance: Struggles in connecting educational research and social justice. In J. Smyth (Ed.), *Qualitative approaches in educational research*, (pp. 43-56).

Adelaide: Flinders University of South Australia.

Smith, J. (1990). Alternative research paradigms and the problem of criteria. In E. Guba (Ed.), *The paradigm dialog*, . Newbury Park: SAGE Publications.

Smith, J. (1993). *After the demise of empiricism: the problem of judging social and education inquiry*. Norwood, N.J.: Ablex Publishing Corporation.

Smyth, J. (Ed.). (1994). *Qualitative approaches in educational research*. Adelaide: Flinders University of South Australia.

Soucek, V. (1993). Is there a need to redress the balance between systems goals and lifeworld-oriented goals in public education in Australia? In C. Collins (Ed.), *Competencies: The competencies debate in Australian education and training*. Canberra: Australian college of Education.

Spearritt, D. (Ed.). (1980). *The improvement of measurement in education and psychology*. Hawthorne, Victoria: Australian Council for Educational Research.

Stake, R. (1991). The countenance of educational evaluation. In M. McLaughlin & D. Phillips (Eds.), *Evaluation and education: at quarter century*, (pp. 67-88). Chicago: the University of Chicago Press.

Stanley, G. (1993). The psychology of competency-based education. In C. Collins (Ed.), *Competencies: the competencies debate in Australian education and training*. Canberra: Australian College of Education.

Stern, D. (1991). *Diary of a baby*. London: Fontana.

Sternberg, R. (1990). T & T is an explosive combination: technology and testing. *Educational Psychologist, 25(3&4)*, 201-222.

Sydenham, P. (1979). *Measuring instruments: tools of knowledge and control*. London: Peter Peregrinus Ltd.

Taylor, C. (1994). Assessment for measurement of standards: The peril and promise of large-scale assessment reform. *American Educational Research Journal, 31(2),* 231-262.

Taylor, P. (1961). *Normative discourse*. Englewood Cliffs: Prentice-Hall, Inc.

The Flinders University of South Australia. (1997). *Calender*. Adelaide: Flinders University of South Australia.

The Social Development Group. (1979). *Developing the classroom group: How to make your class a better place to live in* . Adelaide: South Australian Education Department.

The Social Development Group. (1980). *How to make your classroom a better place to live in* . Adelaide: South Australian Education Department.

Thompson, P., & Pearce, P. (1990). *Testing times*. Adelaide: TAFE National Centre for Research and Development.

Thompson, W. (Ed.). (1987). *Gaia, a way of knowing*. Hudson: Lindisfarne Press.

Travers, E., & Allen, R. (1994). *Random sampling of student folios: a pilot study (10)*. Brisbane: Board of Senior Secondary School Studies, Queensland.

Watzlewich, P. (1974). *Change*. New York: W Norton & Co.

Weiss, C. (1991). Evaluation research in the political context. In M. McLaughlin & D. Phillips (Eds.), *Evaluation and education; at quarter century*, (pp. 211-231). Chicago: The University of Chicago Press.

Wheeler, L. (1993). Reform of Australian vocational education and training: A competency-based system. In c. Collins (Ed.), *Competencies: the competencies debate in Australian education and training*. Canberra: Australian College of Education.

Wiggins, G. (1988). Teaching to the (authentic) test. *Educational Leadership*(September), 41-47.

Wilbur, K. (1977). *The spectrum of consciousness*. Wheaton: Quest.

Wilbur, K. (1982). *Up from Eden: A transpersonal view of human evolution*. Boston: Shambhala.

Wilbur, K. (1991). *Grace and grit*. North Blackburn: Collins Dove.

Wilbur, K. (1995). *Sex, ecology, spirituality*. Boston: Shambhala.

Wiliam, D. (1995). Technical issues in criterion-referenced assessment: evidential and consequential bases. In T. Kellaghan (Ed.), *Admission to higher education*. Dublin: Educational Research Centre.

Williams, F. (Ed.). (1967). *Educational evaluation as feedback and guide*. Chigago: The National Society for the Study of Education.

Willmott, A. S., & Nuttall, D. L. (1975). *The reliability of examinations at 16+*. London: Macmillan Education Ltd.

Wilson, N. (1966). *A programmed course in physics, Form V*. Sydney: Angus and Robertson.

Wilson, N. (1969). Group discourse and test improvement. Unpublished data.

Wilson, N. (1969). A study of test-retest and of marker reliabilities of the 1966 commonwealth secondary scholarship examination. *ACER Information Bulletin, 50(1).*

Wilson, N. (1970). *Objective tests and mathematical learning*. Melbourne: Australian Council for Educational Research.

Wilson, N. (1972). *Assessment in the primary school*. Adelaide: South Australian Education Department.

Wilson, N. (1974). *A framework for assessment in the secondary school* . Adelaide: South Australian Education Department.

Wilson, N. (1985). *Young people's views of our world* (Peace Dossier 13). Melbourne: Victorian Association of Peace Studies.

Wilson, N. (1986). Programmes to reduce violence in schools. Adelaide: South Australian Education Department.

Wilson, N. (1992). *With the best of intentions*. Nairne: Noel Wilson.

Withers, G. (1995). Achieving comparability of school-based assessments in admissions procedures to higher education. In T. Kellaghan (Ed.), *Admission to higher education*. Dublin: Educational Research Centre.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment, *Review of research in education*, (Vol. 17, pp. 31-71).

Wolf, R. M. (1994). The validity and reliability of outcome measures. In A. C. Tuijnman & T. Neville Postlethwaite (Eds.), *Monitoring the standards of education.*

Wood, R. (1987). Aspects of the competence-performance distinction: Educational, psychological and measurement issues. *Curriculum Studies, 19(5),* 409-424.

Wood, R. (1987). *Measurement and assessment in education and psychology*. London: The Falmer Press.

EPAA
• enter the archives • browse the abstracts • the editors • the edit board
• submit article • submit commentary • search • subscribe
volume: • 1 • 2 • 3 • 4 • 5 • 6 • 7

This article has been retrieved 2099 times since June 12, 1998.

Education Policy Analysis Archives

Volume 6 Number 11 **June 12, 1998** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal.
Editor: Gene V Glass Glass@ASU.EDU.
College of Education Arizona State University,Tempe AZ 85287-2411

Copyright 1998, the EDUCATION POLICY ANALYSIS ARCHIVES.Permission is hereby granted to copy any article provided that EDUCATION POLICY ANALYSIS ARCHIVES is credited and copies are not sold.

Public Policy on Distance Learning in Higher Education:
California State and Western Governors Association Initiatives

Gary A. Berg
Chapman University

Abstract

The Western Governors University (WGU) and the California Virtual University (CVU) are revealing examples of the complex issues involved in implementing distance learning on the public policy level. Although technology is certainly important, it has masked the fact that the WGU and CVU initiatives mark the rise of learner-centered higher education and the increased role of business in the academy. In comparing and contrasting WGU and CVU, it is clear that the WGU is a more radical proposition because of competency-based credit and the connection with private industry. Two important issues driving public policy are raised in these two efforts: First, are the California and Western Governors Association initiatives the product of the commercialization of education or the result of a reform of higher education that may lead to an increased learner-centered orientation? Second, what is the appropriate role of private industry in higher education?

Introduction

Distance learning has become the focus of a great deal of attention in higher education circles in the past few years. While a fascination with the technology has led to enthusiasm from many, it has met with equally intense opposition from others. On the policy level, the Western Governors University (WGU) and the California Virtual University (CVU) are revealing examples of the complex issues involved in implementing distance learning. Although the technology is certainly important, it has masked the fact that the WGU and CVU initiatives mark the rise of learner-centered higher education and the increased role of business in the academy.

In comparing and contrasting these two policy efforts the following key issues emerge:

- private industry in higher education
- competency-based vs. seat time credit
- university governance/faculty labor issues
- accreditation
- education vs. training
- state residency and funding
- consumerism in education

I will not address here the learning theory debate about the validity or value of distance learning, but will instead focus on the policy issues, as well as the forces behind the policies. The overall organization of this article is to look first at the recent history of policy efforts in California and through the Western Governors Association, then examine the debate surrounding key issues, and finally draw conclusions which point to future directions in higher education policy.

History of Western Governors University

The official origin of the Western Governors University was a Memorandum of Understanding that followed the positive reception of a report called "From Vision to Reality." The memorandum cited specific needs that it wanted to address including access, affordability, and certification.

> The strength and well-being of our states and the nation depend increasingly on a strong higher education system that helps individuals adapt to our rapidly changing economy and society; and States must look to telecommunications and information technologies to provide greater access and choice to a population that increasingly must have affordable education and training opportunities and the certification of competency throughout their lives (Western Governors Association, 1996).

In the subsequent Resolution 96-002 signed on June 24, 1996, the Western Governors Association also agreed to support collaboration with businesses, between universities, and among states on financial aid issues. The Governors charged a design team with creating a design plan for a virtual university describing how such an entity could be developed and financed. The primary elements of the mission of this entity adapted from the "From Vision to Reality" document were identified as expanding access, formal recognition of skills and knowledge, shifting the focus of education to competence from "seat time," and new approaches to teaching and assessment. The strategic implementation would be based on a market-orientation that is learner-centered, accredited, competency-based, regional and quickly initiated. In their prospectus, the design team identified their basic approach as creating broader markets for existing educational services, fostering the development of new products where unmet needs are identified, utilizing market mechanisms, and removing barriers to interstate flows of educational activities. Further, they identified the role of the WGU as to provide the means for assessing an individual's competence, act as a vehicle for identifying providers of educational programs, and to provide support services.

Most importantly, the prospectus advocated the creation of regional centers franchised by the WGU as points of access for services. These regional centers would not necessarily be existing educational institutions. Organizations will apply to become regional centers and for-profit businesses will not be excluded. The WGU will also contract with providers of educational materials and assessment instruments. Essentially, the WGU is promoting the creation of both a consortium and a new educational institution which is separately accredited. The role of WGU will be to provide centralized governance, policy guidance and quality control.

Currently, the WGU is in a pilot phase. It is forming the administrative staff and has received "Eligibility for Candidacy" status from the Inter-Regional Accreditation Committee (IRAC). Initially it will focus on the offering of A.A. degrees and certificates rather than bachelor's degrees.

History of California Virtual University

In 1989 the California Legislature approved Senate Bill 1202 which directed the California Postsecondary Education Commission (CPEC) to develop a State policy on distance learning. The resulting report, "State Policy on Technology for Distance Learning" suggested a policy emphasizing equity, quality, diversity, efficiency, and accountability. However, largely because of the extreme funding cutbacks in the early nineties, the distance learning plans could not be implemented by the legislature.

In 1996 the economy began to turn around in California and the distance learning initiatives were picked up again. With the projection of 450,000 additional college-age students over the next 10 years in California by CPEC, the legislature looked at technology as a partial solution. CPEC subsequently wrote two reports, "Moving Forward: A Preliminary Discussion of Technology and Transformation in California Higher Education" (CPEC, 1996) and "Coming of Information Age in California Higher Education" (CPEC, 1997), which attempted to address the need for an overall state-wide approach to technology in education. A third report from CPEC which will focus on research in connection with distance learning theory is due to be released late in 1998.

Executive Order W-153-97 established the California Virtual University Design Team with the charge of recommending a blueprint to meet somewhat vague needs.

> ... by which California-based institutions of higher education may serve the needs of California students and employers through emerging technology-enhanced educational programs, as well as reach national and global demand for such programs and content (State of California, 1997).

In the 1998-99 budget, Governor Wilson has requested a total of $14 million to encourage distance learning, with $6.1 million specifically earmarked for CVU. Wilson's plans include $1 million each for UC and CSU to develop online courses, and $3.9 million for the California Community Colleges (Coleman, 1998).

Assembly Bill 2431 was introduced on February 20, 1998 paving the way for creating standards of distance learning practice in California and establishing the Matching Grant Program to assist California institutions in the development of distance learning courses. In the text of the Bill it is stated: "Distance education shall be utilized by the state to achieve its goals for education, equity, quality, choice, efficiency, and accountability (State of California, 1998)."

This Bill also advocates collaboration between the private sector and educational institutions. In relationship to industry involvement, AB2431 says:

> The state shall encourage collaboration between the private sector and the educational institutions in the use of technology both to enhance the quality of education in the classroom and to expand and enhance the delivery of educational services to homes and worksites.

In a separate but related effort called the California Educational Technology Initiative (CETI), the CSU system proposed an agreement with corporate sponsors to provide an infrastructure for distance learning at CSU campuses. Under great criticism by faculty groups and parties concerned with the business ties with CSU, the proposal is being revised and Microsoft has been dropped from the list of partners.

Analytic Comparison of WGU and CVU Efforts

The following chart shows the similarities and differences between the Western Governors Association and California State distance learning policies.

	WGU	**CVU**
Competency-based credit	x	
Inter-State	x	
Learner-oriented	x	
Private Industry Involvment	x	x
Separately Accredited	x	
Brokering Function	x	x
Financial Imperative	x	x
Training Orientation	x	
Hardware Infrastructure		x

The Western Governors Association and California efforts are similar most importantly in their brokering management approach. As an article called "Western Governors U. Takes Shape as a New Model for Higher Education" in *The Chronicle of Higher Education* reveals, the WGU sees itself to some degree as an enormous course broker: "Governor Leavitt, in fact, likens the new institution to 'a kind of New York Stock Exchange of Technology-delivered courses (Blumenstyk, 1998)." Although the WGU is seeking separate accreditation, it remains to be seen if it will develop its own courses to any extent. In this way, the WGU and CVU avoid obvious competitive battles with existing higher education institutions through a brokering mandate. However, this strategy also severely limits the real impact and value of both of these institutions. In looking at the present offerings of both the WGU and CVU, they are not very impressive. In fact, they are little more than a hodgepodge catalogue of previously existing courses with great differences in format and quality. While the number and quality of these courses is likely to improve, without new course development and overall academic planning the curriculum is likely to remain fragmented.

The Western Governors Association and State of California both are encouraging participation from private industry, which has opened them to criticism. Furthermore, the stated objectives of both organizations are similar in their declared aims of meeting changing student and business needs, providing access for the increased student population, and in increasing the quality of distance format courses.

While there are obvious similarities between the Western Governors Association and California State efforts at creating distance learning institutions, there are important differences. Overall, the Western Governors Association effort is both more ambitious and further developed. As reported in the "San Francisco Examiner," the WGU is seeking separate accreditation while CVU will defer to the sponsoring university for credit.

> The major difference between the two proposals is that students in the Western Governors 'distance learning' program would receive credit from the newly created

> 'WGU,' while those studying via the California linkup would receive credit from participating institutions--which may include Stanford University, UC Berkeley, USC and others (Raine,1996).

This necessarily gives California's effort less impact because students will not be able to complete a degree through the virtual university, only through individual institutions. Second, the WGU is a multi-state effort, while the CVU is exclusively based within California. This makes the WGU's implementation much more difficult--and ultimately more important if it is successful--because it will have addressed the serious financial, funding, and transferability issues that go along with interstate cooperation. In addition, the WGU has a training orientation in its initial curriculum and has decided to focus on A.A. degrees at the outset, rather than bachelor degrees. Undoubtedly this decision is a result of the influence of its corporate advisors--in particular Novell with its CNE training. It is difficult to tell if this emphasis on training and on A.A. degrees is a strategic marketing decision or an academic one. If it is an academic policy decision, the WGU has not explicitly excluded more advanced degrees from their plans at this time.

Training versus a traditional education model is clearly a preoccupation for the WGU. Conversely, California's effort focuses to a great extent on building a technological infrastructure for their three enormous higher education systems. While the Virtual University catalogue in California also lists independent institutions, they are left out of the infrastructure plans. As a consortium of various state and private institutions, the WGU has more difficulty addressing infrastructure issues by legislative measures.

Perhaps the single most important difference concerns the issue of competency-based credit. While California State is experimenting with competency-based credit at CSU Monterey, this is not part of the Virtual University planning thus far. For the WGU, competency-based credit is integral to the overall theory and implementation of their distance learning. The wide-spread implementation of competency-based credit would in fact be revolutionary in its affect on higher education administration.

Analysis of Central Issues

Competency-based Credit

Probably the most radical aspect of the entire WGU effort is its promotion of the complete reforming of university credit based on competency not seat time. While there is some precedent for this action in terms of high school diplomas based on comprehensive testing and limited credit for "life experience" at the college level, competency-based credit faces stiff opposition in terms of transferability and financial incentives for institutions. First, how will other academic institutions regard competency-based degrees from WGU in application to graduate programs? If the degrees are not recognized, this is going to severely affect enrollment. Second, how will universities be compensated for the granting of competency credit? If an institution has no financial incentive for the granting of competency credit, they are likely to see this approach as very much against their interests. As the WGU report entitled "The Policy Environment for Implementing The Western Governors University" indicated:

> The success of these new competency-based approaches will depend on changes in the financial incentives for both students and institutions. If the state's four-year institutions recognize that it is in their financial self- interest to emphasize competencies rather than course- specific credit hours in looking at potential student transfers, their attitudes may change regarding students whose competencies have been certified through the WGU (The Western Governors Association, 1996).

Furthermore, the WGU will need to develop very specific guidelines for the granting of competency credit. In "Concept Paper on System for Credentialing," the WGU puts forward basic premises and directions for the credentialing system they are likely to utilize. Their main premises are: 1) developmental-- focus on on-going diagnoses of the student, not just ending testing; 2) Non-exclusionary--open to everyone; 3) Non- punitive--students are given credit for passing certain sections of tests and will not need to retake those parts; 4) Portable-- transferable skills and knowledge which can be used in multiple settings. To any university administrator looking at this list, it would be clear that this kind of credentialing is going to involve a great deal of staff time. On-going diagnoses, non- exclusionary, modular and portable credentialing is likely to be very time-consuming and would change the role of the institution from teaching to assessing in a large way.

In California, higher education is moving much more cautiously into this notion of competency-based credit. CSU Monterey is one of the few institutions experimenting with competency-based credit in which at the end of their studies students are required to demonstrate competency regardless of accumulated credits or seat-time in order to receive a degree.

State Residency and Funding Issues

As they identify themselves in "The Policy Environment for Implementing the Western Governors University," the WGU has many problems to deal with in regard to residency and state funding including financial aid and residency tuition rates.

> The problem is that existing state policies, even in their most fully-developed form, are increasingly inadequate to handle new forms of postsecondary delivery that make state boundaries essentially irrelevant. The issue of physical presence is at the heart of the problem...States are clearly in

> a period of uncertainty about how to address the challenge of educational programs offered through the Internet by providers with no physical presence in the state -- or in some cases within the United States. No clear legal or policy guidance appears to be available (The Western Governors Association, 1996, p. 3-4).

Most importantly, state authorization also plays a critical role in determining institutional eligibility for federal student assistance under Title IV of the Higher Education Act. How is this to be done with courses offered in cyberspace? Would no federal financial aid be available?

California avoids many of the problems that the WGU has by focusing on California residents. However, if the Virtual University draws students from out-of-state to credit courses, they also will have to deal with financial aid and funding issues as well.

Accreditation

The WGU sought separate accreditation and on May 8, 1998 received notification of gaining "Eligibility for Candidacy" status through the Inter-Regional Accreditation Committee (IRAC) (The Western Governors Association, 1998). IRAC was formed through the collaboration of four regional accrediting associations including North Central Association of Colleges and Schools, Northwest Association of Schools and Colleges, and the Western Association of School and Colleges. As a group, the four associations granted IRAC the power to develop an accrediting process for WGU. This represents a real change in accreditation practice and may be the most significant policy evolution to come about from the efforts of WGU. However, serious questions remain for IRAC. Can a consortium of universities without separate faculty have its own accreditation? In addition, how will IRAC deal with competency-based credit?

Financing

State financing policies also are a hurdle for the WGU. Public policies regarding financing of postsecondary education, both federal and state, are usually based on a measure of clock- hours of instruction. In contrast, the WGU will certify learning on the basis of assessment of competencies. Consequently, it is a real question as to how states can allocate resources for the WGU.

Private Industry in Higher Education

For many critics of the use of distance learning in higher education, the issue is not the use of technology but the perceived commercialization of the academy. The strong reaction to The California Educational Technology Initiative (CETI) proposed by CSU to contract with four large technology corporations (Microsoft, GTE, Fujitsu, and Hughes Electronics Corp.) to provide technology and networking to CSU campuses is a current example of this reaction.

The deal was put on hold at the end of 1997 when faced with widespread criticism from students and non-participating companies with complaints about the privatization of CSU as a whole. The agreement has now been delayed until the May, 1998 Board of Trustees meeting. However, the state's legislative counsel, Bion M. Gregory, released a 27- page review of the plan at the end of January, 1998, with the opinion that the deal was illegal because it would put the university in the role of a profit-making entity (Young, 1998). Contrary to this opinion, others defend the agreement because it provides much needed funding for technology infrastructure and allows for open bidding for services and equipment. Furthermore, it is argued that the agreement does allow CSU to go to other providers for a lower price if necessary (Wilson & David, 1998, p. B15).

University Governance

As a Los Angeles Times article suggests, this conflict between the corporate and academic worlds is centered on the issue of university governance.

> Underlying the misgivings of many academics about the trends illustrated by CETI, the THEN and virtual universities is the suspicion that administrators, legislators and university trustees, under pressure because of mounting technology expenses, are capitulating to the high-tech industry's political agenda, which is clearly hostile to educational principles such as faculty governance and social critique. In other words, some academics are starting to view their institutions as emergent clones of market-driven high-tech companies instead of as universities and colleges. Recent attacks on tenure across the country--a principle not coincidentally reviled by many high-tech leaders--only fuel such suspicions (Chapman, 1998, D6).

While there are real academic autonomy and quality issues at stake in this debate about the involvement of business in education, it is becoming increasingly clear that this is at least partially a labor issue. In fact, the strongest critics of CETI in California have been faculty groups. David Noble to some extent voices the viewpoint of some faculty members in seeing new technology as a tool in labor/management struggles: "As in other industries, the technology is being deployed by management primarily to discipline, de-skill, and displace labor (Noble, 1998, p. 7)."

Conclusion--Towards a Policy on Distance Learning in Higher Education

What are the political forces driving these two pieces of higher education public policy? For David Noble, distance learning is driven by business and university administration collaboration seeking profit and control: "a battle between students and professors on one side, and university administrations and companies with ‘educational products’ to sell on the other (Noble, 1998, p.1)." Do students want distance learning, or is it being forced on them by administrators and high tech corporations? In spite of Noble’s argument, it is hard to ignore indicators such as the finding in the 1995 study from Washington State University which stated that "Teaching conducted only in the traditional campus classroom will not meet the public’s demand for tailored educational services (Dillman, Christenson, Salant, Waner, 1995)." Furthermore, when the University of Colorado at Denver began to offer online courses this past year it found that out of 609 enrollments, 500 where also enrolled in regular courses and therefore did not need to take the courses because of geographic distance--they in fact for one reason or another preferred this delivery method (Guernsey, 1998). For Utah Governor Michael O. Leavitt, the people are demanding a virtual university: "This isn’t something that we’re inventing. The market is driving it. People are demanding it (Blumenstyk, 1998)." There are two central questions in this debate: First, are the California and Western Governors Association initiatives the product of the commercialization of education or the result of a reform of higher education which may lead to an increased learner-centered orientation? Second, what is the appropriate role of private industry in higher education?

In attempting to answer these questions, we need to examine the evolving role of higher education in society, the relationship between education and business, and the administrative structure of universities. For David Noble, the adoption of distance learning leads to the commercialization of higher education. For reformers such as Carol Twigg from EDUCOM and a member of the WGU design team, the traditional system is operating under a manufacturing model in which educational products are created and then pushed onto the marketplace regardless of student needs.

> Our institutions are reminiscent of other kinds of industrial age organizations such as the factory and the department store--characterized by size and centralization--in contrast to the distributed, networked organization and mail-order shopping services of the 1990s (Twigg, 1994, p. 4).

In business terminology, what Twigg is advocating is a marketing or pull strategy, rather than a manufacturing or push orientation. However, Noble and Twigg are talking from two completely different frames of reference. Noble sees higher education as being automated by distance learning, while Twigg envisions a redirection of education through technology so that it is more oriented towards student needs.

This automation versus redirection analogy is a revealing one because it is very much at the center of the debate regarding the implementation of distance learning in universities. Many of the issues and problems surrounding distance learning are a result of conceiving of the use of technology to automate traditional teaching. While there is a great deal of evidence and experience to show that distance learning through videotape and the internet can be very successful, conceiving of it as an imitation of the classroom experience leads inevitably to negative comparisons. In contrast to this, Twigg emphasizes an opportunity to employ a constructivist, learner-centered learning approach through the use of technology. In fact, constructivist learning theory is specifically mentioned in both CVU and WGU documents as an advantage of the use of educational technology in higher education. The public preoccupation and fascination with technology has masked the fact that the WGU and CVU initiatives are indicative of a broader debate about faculty-centered versus learner-centered education on both the level of learning theory and management in higher education.

What are the political forces which are putting these policy initiatives in the limelight? The single most important factor is the changing demographics of higher education leading to a great increase in the average age of students in higher education. One-third of all undergraduate-level and two-thirds of master's-level enrollments are part-time. The largest single demographic group among part-time degree students is women over the age of 35 years old (NUCEA, 1994). The Fielding Institute which offers PhDs through an innovative distance format program reports that their average student age is a remarkable 46 years old (WASC Annual Conference, 1998). Increasingly, the needs of the traditional 18-22 year old college student are being overwhelmed by a much larger and more demanding group of adults needing and wanting lifelong educational opportunities. This extreme change in the composition of the student body of the university is having a dramatic affect on higher education, one which is likely to be even more important than the effect of the G.I. Bill in altering the composition of American universities. What this means is that because the students are different their needs are different, and the function of the university is changing along with the new student body. The historical role of liberal arts institutions to provide a broad-based education which prepare young adults to become productive citizens is no longer appropriate for the majority of higher education students. Many of these students are already accomplished professionals with families who often are already actively involved in American society, and more importantly, the political system.

A second force behind public policy in higher education is business and the increased need for a skilled workforce. In *The Monster Under the Bed*, Stan Davis and Jim Botkin argue that we are seeing a transition of higher education from government control to business control as a result of the changing needs of students and the role of education moving increasingly towards preparation for a job. On local, state, national, and international scales, the demand for higher education is pushing universities to become more productive and efficient. Consequently, public policy makers are looking increasingly towards business for answers. Universities are enormously labor-intensive endeavors as presently constructed. Faculty and staff costs make up approximately 80 percent of the budget of colleges and universities (Twigg, 1996, p. 5). It is a frequent complaint that the use of technology in higher education (and business for that matter) has increased expenses instead of lowering them.

Furthermore, distance learning courses usually end up requiring more, not less faculty time. However, distance learning methods still offer the possibility of dealing with the enormous labor expense in a business-like manner by creating educational capital through technology products. As Carol Twigg points out in "Academic Productivity: The Case for Instructional Software, A Report from the Broadmoor Roundtable": ... colleges and universities need to find ways to substitute capital for labor in order to improve productivity (Twigg, 1996, p. 5)." Of course, this attempt to make higher education efficient by reducing labor costs is exactly why faculty members feel threatened. More importantly, it remains to be seen if technology will ever effectively reduce faculty labor expenses.

A third reason for this new "consumerism" in education is the ascendancy of the baby boom generation to political power. A highly educated group--not long removed from the curriculum power struggles of the 60s and 70s and often with college-age children--they are determined to see educational institutions become more responsive. While this generation does have respect for notions of academic freedom and the value of intellectual pursuits, they are suspicious of wasteful bureaucracies. These educational consumers see a great deal of inefficiency in traditional higher education and an alarming lack of attention to undergraduate education.

Some might argue the following: Surely a changing student population isn't reason enough in itself to reformulate what has been a very successful university system in the U.S. Presumably, students still go to school to learn from faculty. Students are not going to teach themselves. Students do not always understand a given academic field well enough even as adults to make good decisions about their own education. What is important here is to distinguish between learner-centered and learner-taught higher education. At a recent WASC annual conference (WASC Annual Conference, 1998) , Carol Twigg responded to a similar challenge by raising the analogy of the doctor- patient relationship. While patients do not want to perform surgery on themselves, in an age of managed-care dominance, they do want hospitals to be more responsive to their needs. The comparison is apt. For the most part, students do not want to teach themselves. However, to take this analogy further, do we want decisions made about our health and education based on the bottom line of a business? Isn't it important that some key areas of human endeavor be protected to some degree from the inevitable excesses of capitalism? Some might argue that non- profit higher education institutions are already dominated by financial decisions, and of course it would be very naïve to believe otherwise. However, non-profit institutions regularly make decisions which benefit and enrich their students based on their overall institutional missions which have nothing whatever to do with the bottom line. It is doubtful that profit-making educational institutions would act similarly. On the other hand, non-profit universities do have a great deal that they can learn from businesses, starting with the marketing principle of staying close to the customer (student). Additionally, the interests of businesses and non-profits do come together in the common goal of creating an educated workforce.

Corporations have moved reluctantly into the business of educating and training their employees; and if higher education institutions become more responsive, businesses will gladly give up this role.

On the public policy level, the California and Western Governors Association initiatives reveal two different kinds of approaches to distance learning. While some might describe the differences as being a centralized versus decentralized kind of opposition, they might better be described as faculty-centered versus learner-centered. The CVU is obviously much more conservative and anchored in the control of the existing educational institutions with its faculty governance schemes. Under this model, technology will be used to augment traditional classroom courses and probably only have widespread use through continuing education, which is historically much more market- driven and flexible. Clearly, with a proposed $6.1 million in the coming fiscal year for the UC, CSU and community college system, the California Virtual University is a small effort. While it is difficult for anyone in the California Legislature to argue against technology in higher education, it is hard to not view this effort as something of a cynical public relations effort with little real consequence. It is unlikely that as presently conceived that the CVU will have much immediate impact on access to degree credit courses in California and will certainly have minimal effect in meeting the increased enrollment projections of Tidal Wave II. As CPEC concluded in its 1996 report "Moving Forward," California needs more aggressive leadership in higher education.

> There appears to be widespread agreement among educational planners working on a regional basis that what California needs is leadership that moves public colleges and universities to a completely new paradigm that is student-centered (California Post-secondary Education Commission, 1996, p. 15).

The California Virtual University clearly does not represent an instance of this kind of leadership.

On the other hand, the WGU effort has higher education reform at its philosophical roots with the insistence on its own accreditation, competency-based credit, and partnerships with businesses with an eye towards training instead of education. Of course, because it is more ambitious, the WGU plans are going to be more difficult to implement. Nevertheless, the WGU is likely to have a greater impact on the future of higher education in the United States.

However, the WGU reliance on competency-based granting of credit is not without philosophical problems. Testing in education in America has reach epidemic levels, from continual preoccupation with the classification of K-12 students to graduate admissions tests. Traditionally, tests are designed to assess what students know, not what they have learned. In this way we are in fact already using a competency-based model in higher education. The difference is that we are also requiring a certain amount of seat-time regardless of ability or non-academic background. I think that the WGU competency-based model puts an unfortunate emphasis on competency instead of the learning process. Because the WGU is placing so much emphasis on a competency model, the assessment instruments used are going to have to be more than behavior-oriented standardized tests. While the WGU planners seem to recognize this in their planning documents, I think that it is going to be very easy to slip into a standardized test model. What is really missing in the assessment emphasis is adequate assessment of incoming students on a course-by-course basis. Certainly faculty members have been assessing students needs on an ongoing basis in the classroom for years. If you remove the immediate contact with the instructor, how can higher education truly address individual needs without adequate up front assessment? Finally, can computer-based programs accurately assess the kind of complex knowledge striven for in higher education? On a practical level, this kind of assessment both incoming and outgoing is likely to be very expensive if it is useful.

On the policy level, the most important new initiative would be one which gives funding to research projects leading to a better understanding of technology-enabled learner-centered education. In some ways the need is presently out-running the knowledge base in terms of the use of technology in the classroom. Some critics feel that institutions are jumping into distance learning before really understanding its value as an approach to learning. While there are thousands of research studies and many years of experience in various kinds of distance learning, there is a certain amount of justification in this viewpoint. Educational technology is still in its infancy. In many ways we are at a stage in education very similar to that of the early film industry, which began by simply recording Broadway stage plays. We are still imitating the classroom with educational technology and consequently offering once-removed imitations of the in-person experience. It was a number of years before film developed its unique language and power as a medium, and educational technology as a new medium faces this same developmental challenge. While we are still developing the language of technology-mediated education, the best use of public funds for distance learning might be in gaining a better understanding of these important new tools through research.

In looking to the future, public policy in relationship to distance learning must address the key issues of credit, transferability, financial aid, and interstate enrollment policies. These are all issues that the WGU is addressing and they are consequently playing an important role in the history of higher education. In terms of leadership at the policy level, the CVU represents a very modest effort to automate the existing faculty-controlled academic institutions. In the final analysis, the marketplace for education is going to be the most important factor, not public policy. If the state and federal government do not respond to the increased demand for learner-centered models of higher education, more flexible independents and for- profit institutions will meet the need. In fact, this is already happening. Nevertheless, I think that it is important on a policy level that the non-profit nature of higher education be protected. Higher education is simply too important on both a personal and social level to leave to the mercy of the free marketplace. The right of citizens to access affordable, quality education must be protected. However, those in non-profit higher education must make the argument that they offer something that the for-profit model will not or cannot. It must prove its value and not simply retreat into a divisive faculty labor stance that the public will view as self-interested. Furthermore, it must pay attention to the learner-centered demands of the public because the needs of the students have changed. Technology can help with this transformation and non-profit higher education would be best served by embracing these new tools rather than engaging in a self-destructive fight in which students will be the big losers.

References

Blumenstyk, Goldie. (February 6, 1998). Utah's Governor Enjoys Role as a Leading Proponent of Distance Learning. *The Chronicle of Higher Education*. February 6, 1998, p. A23.

Blumenstyk, Goldie. (February 6, 1998). Western Governors U. Takes Shape as a New Model for Higher Education. *The Chronicle of Higher Education*. February 6, 1998, p. A21.

California Post-secondary Education Commission. (1997). Coming of Information Age in California Higher Education.

California Post-secondary Education Commission. (1996). Moving Forward: A Preliminary Discussion of Technology and Transformation in California Higher Education.

Chapman, Gary. (January 19, 1998). Will Technology Commercialize Higher Learning. "Los Angeles Times"., p. D1.

Coleman, Donald E. (January 9, 1998). Wilson Pushes Cyber Education in Budget. The Fresno Bee.

Davis, Stan, & Botkin, Jim. (1994). *The Monster Under the Bed*. Touchstone.

Dillman, Christenson, Salant, Waner. (1995). What the Public Wants from Higher Education. SESRC.

Guernsey, Lisa. (March 27, 1998). Colleges Debate the Wisdom of Having On-Campus Students Enroll in On-Line Classes. *Chronicle of Higher Education*.

Noble, David F. (1998). Digital Diploma Mills: The Automation of Higher Education. (Available online: www.firstmonday.dk/issues/issue3_1/noble/index.html).

NUCEA. (1994). *Lifelong Learning Trends*.

Raine, George. (October 3, 1996). California Virtual Would Offer Courses, and Credit, From Many Colleges. "San Francisco Examiner".

State of California Assembly Bill 2431.

State of California. (1997). Executive Department. Executive Order W-153-97.

State of California. (1998). Wilson Announces Corporate Sponsors of California Virtual University. Press Release. January 12, 1998.

Twigg, Carol A. (1994). The Need for a National Learning Infrastructure. *Educom Review*. Volume 29, Numbers 4, 5, 6, 1994.

Twigg, Carol. (1996). Academic Productivity: The Case for Instructional Software: A Report from the Broadmoor Roundtable. July 24-25, 1996.

Twigg, Carol A. and Oblinger, Diana. (1996). The Virtual University: A Report from a Joint Educom/IBM Roundtable. November 5-6, 1996.

WASC Annual Conference. (1998). Newport Beach, CA. April 16, 1998.

The Western Governors Association. (1996). Concept Paper on System for Credentialing.

The Western Governors Association. (1996). Draft Memorandum of Understanding.

The Western Governors Association. (1996). Learner Support Services of the Western Virtual University.

The Western Governors Association. (1996). Resolution 96-002.

The Western Governors Association. (1996). The Policy Environment for Implementing The Western Governors University.

The Western Governors Association. (1996). The Western Governors University: A Proposed Implementation Plan.

The Western Governors Association. (1997). A Proposed Academic Infrastructure for Credentialing at the WGU.

The Western Governors Association. (1998). Western Governors University Granted Eligibility Status. Press Release. May 13, 1998.

Wilson, Blenda & Ernst, David. (1998). Accessing Excellence at CSUN. " Los Angeles Times". January 25, 1998, p. B15.

Young, Jeffrey. (1998). California System Delays Technology Deal

Anew, as State Official Says It's Illegal. *The Chronicle of Higher Education*. February 4, 1998.

About the Author

Gary A. Berg
Director of Extended Education
Chapman University
333 N. Glassell Street
Orange, CA 92866

email address: gberg@chapman.edu

Gary A. Berg is currently Director of Extended Education at Chapman University and has worked in adult education administration for twelve years at Chapman University, the California School of Professional Psychology, the Directors Guild of America, and UCLA Extension. He is involved in various forms of distance learning administration, has written articles on new media, and is currently a doctoral student at Claremont Graduate University.

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

Greg Camilli
Rutgers University

John Covaleskie
Northern Michigan University

Andrew Coulson
a_coulson@msn.com

Alan Davis
University of Colorado, Denver

Sherman Dorn
University of South Florida

Mark E. Fetler
California Commission on Teacher Credentialing

Richard Garlikov
hmwkhelp@scott.net

Thomas F. Green
Syracuse University

Alison I. Griffith
York University

Arlen Gullickson
Western Michigan University

Ernest R. House
University of Colorado

Aimee Howley
Marshall University

Craig B. Howley
Appalachia Educational Laboratory

William Hunter
University of Calgary

Richard M. Jaeger
University of North Carolina--Greensboro

Daniel Kallós
Umeå University

Benjamin Levin
University of Manitoba

Thomas Mauhs-Pugh
Rocky Mountain College

Dewayne Matthews
Western Interstate Commission for Higher Education

William McInerney
Purdue University

Mary P. McKeown
Arizona Board of Regents

Les McLean
University of Toronto

Susan Bobbitt Nolen
University of Washington

Anne L. Pemberton
apembert@pen.k12.va.us

Hugh G. Petrie
SUNY Buffalo

Richard C. Richardson
Arizona State University

Anthony G. Rud Jr.
Purdue University

Dennis Sayers
Ann Leavenworth Center
for Accelerated Learning

Jay D. Scribner
University of Texas at Austin

Michael Scriven
scriven@aol.com

Robert E. Stake
University of Illinois--UC

Robert Stonehill
U.S. Department of Education

Robert T. Stout
Arizona State University

---

archives | abstracts | editors | board | submit | comment | subscribe | search

EPAA
• enter the archives • browse the abstracts • the editors • the edit board
• submit article • submit commentary • search • subscribe
volume: • 1 • 2 • 3 • 4 • 5 • 6 • 7

This article has been retrieved 291 times since July 11, 1998.

Education Policy Analysis Archives

Volume 6 Number 12 July 11, 1998 ISSN 1068-2341

A peer-reviewed scholarly electronic journal.
Editor: Gene V Glass Glass@ASU.EDU.
College of Education
Arizona State University,Tempe AZ
85287-2411

Copyright 1998, the EDUCATION POLICY ANALYSIS ARCHIVES.Permission is hereby granted to copy any article provided that EDUCATION POLICY ANALYSIS ARCHIVES is credited and copies are not sold.

Counseling in Turkey:
Current Status and Future Challenges

Suleyman Dogan

Gazi College of Education
Gazi University
Ankara, Turkey

Abstract In this article a special emphasis is placed on the current status and the future challenges of counseling in Turkey. A brief history of counseling in Turkey, current developments, and the basic issues in this field are pointed out. Finally, the future challenges and recommendations to improve the current status of counseling are discussed.

Counseling in Turkey: Current Status and Future Challenges

The Turkish Republic was built on the ashes of the Ottoman Empire and started her existence under the able leadership of Ataturk in 1923. Turkey is a country of over 63 million people. Situated in both Europe and Asia and controlling the major waterway between the Black Sea and the Aegean and Mediterranean Seas, the country has long been a crossroads between east and west and north and south. A developing country, Turkey is subjected to the usual problems of industrialization and urbanization, including a significant increase in the breakdown of family networks, and the modification of traditional of cultural patterns (McWhirter, 1983).

Turkey has already made major efforts in the education development, and has set ambitious goals in education and training as a paret of its strategy for economic adjustment and national development. The objective of national education is to train "good people," "good citizens," and "qualified manpower," which thus carries implications for counseling in education (Organization for Economic Cooperation and Development, 1989).

The practice of counseling has become international in scope. Counseling is not only emerging in developed countries but also in developing countries, as industrialization replaces traditional paradigms of decision making and career selection (McWhirter, 1988). It is being interpreted differently in different countries, with its form and expression influenced by political, sociological, and economic considerations.

This article aims to provide readers with an overview of the current status of counseling in Turkey in order to enhance awareness of how specific cultural, political, and economic factors have impacted counseling and its delivery systems and how international perspectives have influenced the counseling profession. It may also enable counselors to develop a better perspective of current issues in the counseling profession throughout the world.

The Emergence of Counseling in Turkey

Earlier historical forces that created a need for counseling in the United Sates are currently active in Turkey. Psychological testing, vocational career choice, and mental health concerns are currently paramount in defining what counseling is and what counselors do (McWhirter, 1983).

As in the United States and most other countries, counseling in Turkey began in the schools (Kepceoglu, 1986). Turkey has been making efforts to develop a system of counseling in schools for about fifty years. The first school counselors were primarily teachers, and counseling was a function they performed in addition to their teaching responsibilities. Their major duties involved career and educational counseling.

The main factors that have influenced the emergence of counseling in education settings in Turkey may be listed as follows:

1. Social changes, such as modernization, technological development, democratization, and changing family patterns, have created the need and the desire for counseling in education (McWhirter, 1983).
2. Counseling has been viewed as an effective means for developing human potential (Kepceoglu, 1994).
3. The individual differences emphasized in education have contributed to the emergence of counseling in the schools

(Organization for Economic Cooperation and Development, 1989).

4. The counseling services have been seen as a useful means of modernizing and democratizing the school system (T. C MEB Turkiye Egitim Milli Komisyonu Raporu, 1960).

A Brief History of Counseling in Turkey

The Turkish counseling movement dates back to 1950 and derives largely from advances and developments in the United States system of counseling, such as Rogers's person-centered approach. The history of counseling in Turkey may be organized around five identifiable historical periods. To better understand the development of counseling movement in Turkey, it is necessary to review each period and reflect on the significant events from 1950 to the present.

Taking Initial Steps (1950-1956)

This period between 1950 and 1956 marks the beginning of the counseling movement in Turkey and is considered as the most active period in terms of developments in the counseling field (Kuzgun, 1991; Tan, 1986). The visits of some American counselor educators and the efforts of pioneer counselor educators trained in the United States played a significant role in bringing counseling concepts into Turkish education in the 1950s (Baymur, 1980; Tan, 1986).

This period witnessed the several important developments in counseling. The Ministry of National Education set up a Test and Research Bureau to standardize IQ tests as well as personality and achievement tests for diagnostic and educational purposes in 1953 (Dogan, 1996). A Center for Psychological Services was established in Ankara in 1955. It was later changed to a Guidance and Research Center and exported to other provinces (Oner, 1977). About 100 such centers have been created and 606 counselors called "guidance teachers" were employed in these centers by the 1997-1998 academic year. Neither the quality nor the number of such centers is adequate to meet the demand. These centers focus on "correction" and "remedial" functions, and cater primarily to students in need of special education (Education in Turkey, 1995). The centers cannot reach a large part of the student population, either rural or urban, and the majority of the students, teachers, and parents are uninformed about their purpose and function.

Formative Years (1957-1969)

The second period of counseling, from approximately 1957 to 1969, was the formative years. Although this period is considered an inactive stage for counseling, it witnessed significant events that played an important role in the movement of counseling in Turkey. This period included the adaptation and development of some group tests and rapid changes in school curricula (Oner, 1977). Delivering counseling in secondary schools, especially career counseling, was considered to accomplish the training of qualified manpower through the school years and to solve the unemployment problem as a long-term objective in the 1960s (Tan, 1986). Counseling was proposed as a means of enhancing pupils' well-being by the Seventh Council of National Education in 1962 (Baymur, 1980). Some Turkish universities began to set up either undergraduate or graduate counselor education programs based on the counselor education models of the United States, such as Rogers's person-centered approach (Kuzgun, 1993; Ozguven, 1990).

Establishing Counseling Services in Schools (1970-1981)

In many ways this period was a golden era, very active for school counseling. It marked the beginning of professional counseling practice in schools and witnessed many developments in the counseling field.

The Ministry of National Education implemented some essential policies and then employed 90 counselors to start services for 24 selected secondary schools in the 1970-1971 academic year. Although the number of schools having counseling services and the number of counselors employed increased every year, the rate of increase was very slow (Kepceoglu, 1986). (About 2,199 school counselors were working in 2,033 schools, mostly secondary schools, accommodating 12 million students in Turkey by the 1997- 1998 academic year.) The Ministry of National Education directly addressed counseling (guidance) in the schools in its official documents. These years of the Turkish counseling movement witnessed guidance programs in the secondary schools consisting of extracurricular activities during students' homeroom hours. Homeroom teachers assumed the duties of educating students through various group guidance activities and undertook the responsibility for conducting counseling in schools. The Basic Law of National Education, which was enacted in 1973 and updated in 1983, accepted orientation as a basic principle to be accorded through the education system. Orientation and evaluation of the success attained are to be effected via objective evaluation, test and measurement methods, and guidance services (Education in Turkey, 1995). Both the Tenth National Education Council in 1981 and the Eleventh National Education Council in 1982 focused on the need for counselors in education settings and the establishment of counselor education programs at three different levels in universities. These councils also recommended "counselor" as a title for the graduates and confirmed "guidance" as a specialty field in education (Dogan, 1990).

Establishing Undergraduate Programs in Counseling (1982-1995)

During this period the number of four-year undergraduate counselor education programs rapidly increased. In 1982, the Turkish universities began admitting students to a four-year bachelor of education program with a major in guidance and counseling. (About 19 universities currently offer four-year undergraduate programs, which primarily emphasize school counseling.) The number of masters and doctoral degree programs was also increased. As in the United States, counseling is flourishing in Turkish colleges of education rather than in departments of psychology (Whiteley, 1984). Each university was obliged to establish a counseling and guidance center to be in charge of individual, educational, and career counseling of students as a division of the medical-social, health, culture and sports activities department by the new Higher Education Law in 1984 (Resmi Gazete, No. 18301, 1984).

The Psychological Counseling and Guidance Association was founded in 1989 by a group of counselor educators at Hacettepe University in Ankara. Its current membership of almost 450 is made up mainly of school counselors and counselor educators. The association began to publish the respected *Journal of Psychological Counseling and Guidance* in 1990 and added a newsletter entitled "Psychological Counseling and Guidance Bulletin" in 1997.

The Psychological Counseling and Guidance Association held The First National Psychological Counseling and Guidance Congress in 1991 with subsequent national congresses held every two years. In 1995, the association designated the ethical standards of the counseling profession in order to heighten the standards of the profession (Psikolojik Danisma ve Rehberlik Dernegi, 1995). Three different but complementary counselor roles were identified in the ethical guidelines: the remedial, the preventive, and the developmental.

Assigning Counselors to Schools (1996-The Present)

A rapid increase in the appointment of counselors has taken place during the last few years. The Ministry of National Education has also begun to assign counselors trained specificly as counselors to both elementary and secondary schools. Elementary school counseling was emphasized and reviewed in the Fifteenth National Education Council in 1996. Sponsored by the collaboration of the Higher Education Council and the World Bank, the National Education Development Project for Pre-Service Teacher Education further developed guidance and counseling four-year undergraduate, and master's degree programs in 1996 (YOK/World Bank National Education Development Project Pre-Service Teacher Education, 1996). The government was determined to realize the total implementation of reforms in elementary education and extend national obligatory education from grades 1 through 5 to grades 1 through 8 in 1997. Some new steps are being taken to establish a new structure for elementary schools, including counseling.

The Current Status of Counseling in Turkey

The history of counseling in Turkey is closely related to the history of educational practice and problems in the schools. As each country has its own unique historical background, political system, and economic conditions, any counseling model being appropriated from one society into another will naturally be affected by these factors.

Counseling in Turkey is generally perceived as: 1) a corrective and remedial instrument (Kuzgun, 1991), 2) a means of orienting students toward the schools and regulating the manpower that the country requires (Tan, 1986), 3) a means of disciplining and controlling the students in schools (Dogan, 1995), 4) a means of special education, and 5) homeroom and various educational activities being carried out by ordinary teachers (Dogan, 1991).

Current Developments

Counseling is a new phenomenon in Turkey and there is a lack of information, knowledge and understanding of counseling among the public. The term "counseling" is only used by a small group of professionals. The more familiar term is "guidance" which has connotations of leading, directing, coaching, and advising the students at problematic times (Demir & Aydin, 1996).

The current status of the counseling field in Turkey may be briefly characterized as: 1) working primarily with normal individuals, 2) specializing in the interpretation of standardized tests, particularly group tests, 3) including the field of educational, vocational, and personal adjustment, 4) serving as a source of referrals to specialists in other related areas, and 5) being Rogerian in orientation.

Counseling in Turkey had been very much influenced by developments in the United States. The major models and theories adopted have accordingly been culture bound, being been developed in the main for the white middle/upper classes in a different context. As Skovholt (1988) contended, both Rogerian ideas and standardized tests procedures have been imported from the United States in a way that is not completely positive. It is important to note that in the United States as well, applications of traditional models of counseling are being questioned in programs which respond to the needs of people whose cultural background is not white middle/upper class. With the increased migration of people globally, it is important that the skills and techniques of counseling be modified appropriately to work in other countries and cultures as well.

The concrete results of 50 years of counseling developments in Turkey are: 1) the establishment of guidance and research centers in each province, 2) the establishment of counseling services in some elementary and secondary schools, 3) undergraduate and graduate counselor education programs in universities, 4) "guidance" as an elective course included in the teaching knowledge certification program, 5) some in-service counseling training programs arranged by the Ministry of National Education, 6) counseling units as divisions of the medical-social, health, culture and sport activities department in universities, 7) the Psychological Counseling and Guidance Association, 8) the national psychological counseling and guidance congresses held by the association every two years, 9) the *Journal of Psychological Counseling and Guidance* published by the association, 10) the guidelines of ethical standards, and 11) some tests and textbooks in counseling.

Current Issues

Counseling in the accepted American sense is still very limited in Turkey. The counseling profession in Turkey has not been fully successful to date in spite of persistent efforts. Counseling is less developed, less organized, and still in search of its professional identity. Counseling, and in particular school counseling, is evolving very slowly because of failure to place it in the mainstream of school curriculum in Turkey. Neither the quality nor the amount of such services is adequate to meet the demand. Understanding of the nature and mode of delivery of effective counseling services in Turkey is lacking.

As a profession, counseling is still vaguely identified and confused with other disciplines such as psychology, social work, and even psychiatry. Counseling faces numerous obstacles and limitations in Turkey. Counseling in Turkey does not emphasize the development of the individual's potential nor does it require the receptivity of the person's thoughts and feelings. It is more or less guidance based primarily on the provision of information in a directive and advisory manner.

Both undergraduate and graduate counselor education programs have been increasing in number since 1982. There is great disparity in both the classes offered and the content of the courses from one university to another (Akkoyun, 1995). There is also a lack of standardized selection criteria upon which counseling students can be admitted to counselor education programs. There are no formally recognized requirements for certification as a professional counselor; neither are there procedures for, and official accreditation, of undergraduate and graduate training programs and an agreed upon specialty title and definition in counseling. The counseling field is seriously lacking textbooks and sufficient literature to be used in both undergraduate and graduate programs. There is still great dependence upon American literature and research in this field. The dearth of Turkish literature has reduced the quality of the education programs.

All counseling in Turkey is done under the auspices of the government; there is no private counseling practice.

The general style of education still overemphasizes cognitive learning and school achievement and neglects affective development at the secondary level in Turkey. Counseling is not perceived as being essential in such an educational system which does not recognize the concept of individual differences. Counseling was not instituted from the beginning as a separate and powerful department within the Ministry of National Education Organization, but as a supplementary unit of the special education department. School counseling services have not been presented as an integral part of the education process as defined by the curriculum. Some of the appointed counselors are specialized in different disciplines other than counseling, such as sociology, psychology, education and philosophy. Appointing unqualified graduates to school counseling services and guidance and research centers has caused the misunderstanding that counseling is ineffective and even not useful.

Counseling has a decidedly clinical flavor and the ratio of students to counselors appears to be typically about 4,500: 1. The delivery model is neither developmental in nature nor designed for all students. The students perceive school counseling activities as boring, cumbersome, unnecessary, and ineffective (Demir & Aydin, 1996). The school counselor is usually expected to do some tests in addition to school counseling duties.

The lack of standardized counseling tools such as interest, aptitude, intelligence, and personality tests and the unavailability of organized occupational information causes a great difficulty in the work of counselors. This issue has decreased the quality of counseling services both in schools, and in guidance and research centers.

According to the laws or by-laws regulating counseling services both in schools, and in guidance and research centers, counseling is considered as supplementary service of special education and a means of controlling and disciplining the students (Dogan, 1995). The occupational title of counselor is new and sometimes confusing to those who do not know or agree that counseling is an important "third force" in a school, next to the administration and teachers. There is still a certain resistance against the concept of counseling in schools; many teachers and administrators do not see counseling as an important discipline and think of it as a luxury. It is commonly held that counseling in schools can be offered by ordinary teachers rather than counselors since for years teachers have largely been charged with counseling duties.

There is a misunderstanding among the teachers and the administrators that school counselors are incompetent. Most of the counselors limit themselves to individual counseling and neglect all other guidance services, which gives the impression that such counselors do not perform their jobs adequately. School counselors tend to isolate themselves from the normal flow of school life, expressing an air of importance and a feeling that they occupy a higher status than the teachers. "Guidance" has either been included in the teaching knowledge certification program as an compulsory course or an elective course or completely excluded from this program from time to time. This issue has prevented school principals and teachers from developing a common and sufficient understanding about counseling concepts and practices during their pre-service training.

Future Challenges and Recommendations

Although it is generally considered that counseling will continue to grow and gain a broad base of acceptance and support in Turkey, there are a number of immediate challenges confronting counseling in the country. These include the following:

1. Presenting to the broader society its basic mission and the services which it can deliver to clients,
2. Regaining involvement in the field of prevention,
3. Generating more quality scholarly accomplishments,

4. Anticipating the consequences for client needs of pervasive shifts occurring in the economic structure of society,
5. Being recognized as professionals and having a type of certification similar to that offered by the National Board of Counselor Certification in the United States or the Canadian Counselor Certification,
6. Being accepted in community development centers where individual, group and family counseling can be offered.

During the nearly 50 years that have passed since its beginning in Turkey, counseling has emerged as the profession primarily responsible for planning, interpreting, and delivering counseling to students in grades 9-12.

The following recommendations may ameliorate problematic issues in counseling in Turkey:

1. Counseling should be expanded into non-educational settings, such as correctional institutions, mental health, social work, and rehabilitation facilities in order to promote its professional identity.
2. Training and accreditation standards for counseling programs and practices should be designated in order to gain a professional identity and obtain a legitimate role among other mental health professionals.
3. The number of universities offering degrees in counseling should be reduced by the Council of Higher Education. The contents and the standards of the current counselor education programs should be developed and heightened by the qualified faculty and the use of excellent text books, and other media.
4. The counseling unit should be separated from being a division of the Special Education Department and should be instituted as a separate and powerful department responsible for policy within the Ministry of National Education.
5. An institute should be established and charged with producing tests and other counseling materials.
6. The laws and by-laws regulating counseling services both in schools and in guidance and research centers should be revised according to contemporary counseling principles and concepts.
7. School counseling should be extended to cover both elementary as well as secondary schools. School counseling programs should comprise much more than individual counseling.
8. Each school should have a qualified and powerful counseling unit supplied with enough qualified counselors and all the required tools, materials, and tests. Unless this unit provides qualified services, the importance of counseling for the students may not be understood by parents, principals, and teachers.
9. Counseling activities should be included within the school curriculum for two hours per week and these activities should be carried out as group guidance activities by the counselors. Otherwise, counseling will be perceived as an emergency service intervening in problems only after they occur.

Conclusion

Counseling in Turkey is seen as having an integral role in the educational process, through fostering the development and integration of an individual's many potentials. However, its place in Turkish schools has often caused it to be the target of criticism for regulating both the manpower that the country requires, and youth behavior, over which it has had little control. Most counselors still see students with special difficulties on an individual basis, their contacts being remedial or crisis-oriented. The next decade should give rise to the implementation of the developmental and preventive emphasis by counseling practitioners.

As democratization, industrialization, and urbanization continue to supplant traditional cultural models of decision making, career selection, and human service delivery, the awareness of counseling needs is growing in Turkey. As Turkey looks to the future, there seems to be some justification for optimism. Research, publication, and training programs in counseling have advanced and will continue to advance the counseling movement in a way that is necessary in a modern, democratic, and humane society.

References

Akkoyun, F. (1995). Psikolojik danisma ve rehberlikte unvan ve program sorunu: Bir inceleme ve oneriler [The problem of the relationship between the job title and the training programs in psychological counseling and guidance: A review and recommendations]. *Psikolojik Danisma ve Rehberlik Dergisi*, 2(6), 1-28.

Baymur, F. (1980). Turkiye'de rehberlik calismalarinin baslangici, gelisimi, ve bugunku sorunlari [The beginning, the development, and the current issues of guidance activities in Turkey]. In N. Karasar (Ed.), *Egitimde Rehberlik Arastirmalari* (pp. 3-7). Ankara: Ankara Universitesi Egitim Arastirmalari Merkezi Yayini.

Demir A., & Aydin, G. (1996). Student counselling in Turkish universities. *International Journal for the Advancement of Counselling*, 18(4), 287-302.

Directorate General Press & Information of the Turkish Republic. (1995). Education in Turkey. Ankara: Kurtulus Ofset.

Dogan, S. (1990). Turkiye'de rehberlik kavrami ve uygulamalarinin gelisiminde milli egitim suralarinin rolu [The role of national education councils in development of guidance concept and practices in Turkey]. *Psikolojik Danisma ve Rehberlik* Dergisi, 1(1), 45-55.

Dogan, S. (1991). Baslangicindan bugune Turk resmi dokumanlarinda rehberlik kavrami ve anlayisi: Bir inceleme [The conception and understanding of guidance in Turkish legal documents through the years]. *Psikolojik Danisma ve Rehberlik Dergisi*, 1(2), 29-44.

Dogan, S. (1995). Psikolojik danisma ve rehberlik hizmetleri okullarda disiplin isleri ile kesinlikle karistirilmamalidir [Counseling and guidance services in schools must definitely not get confused with discipline treatments]. *Ogretmen Dunyasi*, 184, 14-16.

Dogan, S. (1996). Turkiye'de psikolojik danisma ve rehberlik alaninda meslek kimliginin gelisimi ve bazi sorunlar [The development of professional identity in the counseling and guidance field in Turkey and related problems]. *Psikolojik Danisma ve Rehberlik Dergisi*, 1(7), 32-44.

Kepceoglu, M. (1986). Some negative barriers to effective counseling and guidance in Turkish schools. *Psychological Reports*, 59, 517-518.

Kepceoglu, M. (1994). *Psikolojik danisma ve rehberlik* [Psychological counseling and guidance] (8th ed.). Ankara: Ozerler Matbaasi.

Kuzgun, Y. (1991). *Rehberlik ve psikolojik danisma* [Guidance and psychological counseling] (2nd ed.). Ankara: OSYM Yayinlari.

Kuzgun, Y. (1993). Turk egitim sisteminde rehberlik ve psikolojik danisma [Guidance and counseling in Turkish education system]. *Egitim Dergisi*, 6, 3-8.

McWhirter, J. J. (1983). Cultural factors in guidance and counseling in Turkey: The experience of a Fulbright family. *The Personnel and Guidance Journal*, 61(8), 504-507.

McWhirter, J. J. (1988). Implications of the Fulbright senior scholar program for counseling psychology. *The Counseling Psychologist*, 16(2), 307-310.

Oner, N. (1977). Psychology in the schools in international perspective, 2. (International School Psychology). Columbus, OH: U. S. Department of Health Education & Welfare National Institute of Education. (ERIC Document Reproduction Service No. ED 147 257).

Organization for Economic Cooperation and Development. (1989). Turkey: Reviews of national policies for education. Paris. (ERIC Document Reproduction Service No. ED 312 193).

Ozguven, E. (1990). Ulkemizde psikolojik danisma ve rehberlik faaliyetlerinin dunu ve bugunu [The past and the present of psychological counseling and guidance activities in our country]. *Psikolojik Danisma ve Rehberlik Dergisi*, 1(1), 4-15.

Psikolojik Danisma ve Rehberlik Dernegi. (1995). Psikolojik danisma ve rehberlik alaninda calisanlar icin etik kurallar [The ethical standards for counselors]. Ankara: 72 TDFO Ltd., Sti.

Skovholt, T. M. (1988). Searching for reality. *The Counseling Psychologist*, 16(2), 282-287.

Tan, H. (1986). *Psikolojik danisma ve rehberlik* [Psychological counseling and guidance]. Istanbul: Milli Egitim Basimevi.

T. C. MEB. (1960). Turkiye egitim milli komisyonu raporu [The report of Turkish national education committee]. Ankara: Milli Egitim Basimevi.

Turkiye Cumhuriyeti. (1984). Resmi gazete [The official gazette]. No. 18301.

YOK/World Bank National Education Development Project Pre-Service Teacher Education. (1996). Teacher education: Guidance and counseling. Ankara: YOK.

Whiteley, J. M. (1984). Counseling psychology: A historical perspective [Special Issue]. *The Counseling Psychologist*, 12(1), 1-126.

About the Author

Suleyman Dogan

Associate Professor
Psychological Counseling and Guidance Program
Educational Sciences Department
Gazi College of Education
Gazi University
Ankara, TURKEY

Phone: 011 90 (312) 2126470/3787
Fax: 011 90 (312) 2238693

Suleyman Dogan received his master's degree and Ph.D. in Psychological Counseling and Guidance from Hacettepe University at Ankara in Turkey. He is a former junior high school teacher, research assistant, assistant professor, and currently an Associate Professor at Gazi University at Ankara, where he trains counselors and teachers. His responsibilities include teaching, student advising, research, and conducting groups. He is the author of several articles and presentations, especially on the development of counseling and related issues in Turkey. He served on the Psychological Counseling and Guidance Association Administrative Committee in Turkey for two years. Dr. Dogan spent his sabbatical leave in Counseling and Higher Education Department at the College of Education, Ohio University during the 1997-1998 academic year.

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

Greg Camilli
Rutgers University

John Covaleskie
Northern Michigan University

Andrew Coulson
a_coulson@msn.com

Alan Davis
University of Colorado, Denver

Sherman Dorn
University of South Florida

Mark E. Fetler
California Commission on Teacher Credentialing

Richard Garlikov
hmwkhelp@scott.net

Thomas F. Green
Syracuse University

Alison I. Griffith
York University

Arlen Gullickson
Western Michigan University

Ernest R. House
University of Colorado

Aimee Howley
Marshall University

Craig B. Howley
Appalachia Educational Laboratory

William Hunter
University of Calgary

Richard M. Jaeger
University of North Carolina--Greensboro

Daniel Kallós
Umeå University

Benjamin Levin
University of Manitoba

Thomas Mauhs-Pugh
Rocky Mountain College

Dewayne Matthews
Western Interstate Commission for Higher Education

William McInerney
Purdue University

Mary P. McKeown
Arizona Board of Regents

Les McLean
University of Toronto

Susan Bobbitt Nolen
University of Washington

Anne L. Pemberton
apembert@pen.k12.va.us

Hugh G. Petrie
SUNY Buffalo

Richard C. Richardson
Arizona State University

Anthony G. Rud Jr.
Purdue University

Dennis Sayers
Ann Leavenworth Center
for Accelerated Learning

Jay D. Scribner
University of Texas at Austin

Michael Scriven
scriven@aol.com

Robert E. Stake
University of Illinois--UC

Robert Stonehill
U.S. Department of Education

Robert T. Stout
Arizona State University

archives | abstracts | editors | board | submit | comment | subscribe | search

EPAA
• enter the archives • browse the abstracts • the editors • the edit board
• submit article • submit commentary • search • subscribe
volume: • 1 • 2 • 3 • 4 • 5 • 6 • 7

This article has been retrieved 2287 times since July 14, 1998.

Education Policy Analysis Archives

Volume 6 Number 13 **July 14, 1998** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal.
Editor: Gene V Glass Glass@ASU.EDU.
College of Education
Arizona State University,Tempe AZ
85287-2411

Copyright 1998, the EDUCATION POLICY ANALYSIS ARCHIVES.Permission is hereby granted to copy any article provided that EDUCATION POLICY ANALYSIS ARCHIVES is credited and copies are not sold.

Consequences of Assessment: What is the Evidence?

William A. Mehrens
Michigan State University

Abstract Attention is here directed toward the prevalence of large scale assessments (focusing primarily on state assessments). I examine the purposes of these assessment programs; enumerate both potential dangers and benefits of such assessments; investigate what the research evidence says about assessment consequences (including a discussion of the quality of the evidence); discuss how to evaluate whether the consequences are good or bad; present some ideas about what variables may influence the probabilities for good or bad consequences; and present some tentative conclusions about the whole issue of the consequences of assessment and the amount of evidence available and needed.

I. INTRODUCTION

It is a pleasure to address friends, colleagues, and associates on what I believe to be an important topic -- what evidence do we have regarding the consequences of assessment. I actually chose this topic at last year's (1997) convention when I attended a symposium on consequential validity. As most of you probably know, I am not a fan of the *term* "consequential validity." However, I am interested in the consequences of assessment, and I hope all of you are also. Last year's symposium had such illustrious speakers as Ross Green, Suzanne Lane, Bob Linn, Pam Moss, Mark Reckase, and Elizabeth Taleporos. It was a great session. While they agreed on many things, I perceived some differences in opinion about the amount, quality and interpretation of the evidence regarding the consequences of assessment. I left that session believing that not enough evidence was available but that it would be worthwhile to review the evidence more thoroughly. Then, last summer (1997), at the Council of Chief State School Officers Large Scale Assessment Conference, Peter Behuniak, Bob Linn, David Miller, and Gloria Turner presented evidence they had regarding the consequences of assessment. While I was very impressed with their scholarship, I again was left feeling it would be worthwhile to investigate the topic further. In addition to the fact that the scholarly presentations mentioned above left me unsatisfied with respect to the evidence on consequences, there are additional rationales for choosing this topic.

Many, but certainly not all, political leaders at the national, state and local levels have been touting the value of large scale assessment. For example, President Clinton and Secretary of Education Riley have argued that voluntary national tests of reading at grade four and mathematics at grade eight would have positive consequences for education. Secretary Riley has said, "I believe these tests are absolutely essential for the future of American education" (Riley, 1997a, quoted from Jones, 1997, p.3) (Note 2). President Clinton has also asked each state to adopt tough standards for achievement. The argument seems to be that if tough standards are adopted, achievement will rise.

Educational reformers suggest that:

> Assessments play a pivotal role in standards-led reform, by: communicating the goals. ... providing targets..., and shaping the performance of educators and students. Coupled with appropriate incentives and/or sanctions--external or self-directed--assessments can motivate students to learn better, teachers to teach better, and schools to be more educationally effective. (Linn and Herman, 1997, iii).

Note the word "can" in the above quotation. The question is, do they? Linn, Baker and Dunbar (1991) pointed out that it cannot just be assumed that a more "authentic" assessment will result in better classroom activities. Linn also correctly suggested that:

> Evidence is also needed that the uses and interpretations are contributing to enhanced student achievement and at the same time, not producing unintended negative outcomes (1994, p. 8).

There is no question that assessment is perceived by many as having a potential for good in both the evaluation of the schools and, if believed necessary, the reforming of them. But there are reasons to question whether that potential will be realized. As Goodling has stated,

> If testing is the answer to our educational problems, it would have

> solved them a long time ago. American students are tested, tested, tested, and the Clinton administration is proposing to test our children again (August 13, 1997).

Goodling suggested that thinking that new tests will lead to better students is "akin to claiming that better speedometers make for faster cars." (quoted from Froomkin, 1997).

There are both potential values and potential dangers in large scale testing. Is the potential value of assessment a vision or an illusion? Are the potential dangers likely to be realized? What is the evidence?

For testing to be a good thing, the positive consequences must outweigh the negative consequences -- by some factor greater than the costs. The costs of large scale assessments are particularly high for alternative forms of assessment such as have been used in Kentucky. Are the consequences of assessment worth the cost? Are the consequences of alternative assessments worth the much greater cost? What is the evidence?

General Overview

In this presentation I wish to spend a brief amount of time on the prevalence of large scale assessments (focusing primarily on state assessments); discuss the purported purposes of these assessment programs; enumerate some potential dangers and potential benefits of such assessments; investigate what the research evidence says (and does not say) about assessment consequences (including a discussion of the quality of the evidence); discuss how to evaluate whether the consequences are good or bad; present some ideas about what variables may influence the probabilities for good or bad consequences; and present some tentative conclusions about the whole issue of the consequences of assessment and the amount of evidence available and needed.

Because the evidence is insufficient, my tentative conclusions about the consequences of assessment will, at times, obviously and necessarily be based on less than adequate evidence. It may seem drawing such conclusions runs counter to a general value of educational researchers -- that inferences should be based on evidence. I am firmly on the side that more evidence is needed and that inferences should be drawn from such evidence. I am firmly against passing pure proselytizing off as if it is research based. I am not opposed to drawing tentative conclusions from less than perfect data. But crossing the line from basing inferences on evidence to basing inferences on a will to believe should not be done surreptitiously, and I try hard to avoid that in this paper.

For those of you who do not make it to the end, I will tell you now that the conclusion will be that the evidence is reasonably scarce (at least insufficient), and equivocal.

II. POPULARITY, PREVALENCE, PURPOSES AND FORMAT OF LARGE SCALE ASSESSMENT PROGRAMS

A. Popularity

Large scale assessment programs are, at the abstract level, popular with politicians and the public. This is true for both proposed and actual state level assessments and the proposed national level "voluntary" assessments. One example of the popularity is obtained from the 29th Gallup Poll. That poll showed that 57% of the public favor President Clinton's proposed voluntary national test (Rose, et al., 1997). However, when proposals get more specific, there can be opposition -- as witnessed by the opposition of many groups after the proposed testing plan got more specific (e.g., with respect to what languages the test would be administered in). It should be noted that front line educators -- those that might be most informed about the potential value of the proposed voluntary national test are far less favorable than the general public. Langdon (1997), reporting on the PDK poll of teachers found that 69% are opposed to Clinton's proposal. Measurement specialists might also have a reasonable claim to being more informed than the politicians or the public. The comments that appeared on the Division D listserve (http://aera.net/resource/) suggest that there are far more negative views among listserve authors about the value of such tests than there are positive views.

B. Prevalence

Regarding prevalence, state programs have been prevalent for at least fourteen years. As early as 1984, Frank Womer stated that "clearly the action in testing and assessment is in state departments of education" (Womer, 1984, p. 3). In identifying reasons for this, Womer stated that:

> Lay persons and legislators who control education see testing-assessment as a panacea for solving our concerns about excellence in education (Womer, 1984, p.3).

Anderson and Pipho reported that in 1984, 40 states were actively pursuing some form of minimum competence testing (1984).

Of course, it turned out that these minimum competency tests were not a "panacea" for concerns about educational excellence, although there exists some debate about whether they were, in general, a positive or negative force in education. In general, there has been a change from testing what were called minimum competencies to testing what we might now call world class standards. And, there has been a bit of a change in how we assess -- with a movement away from sole reliance on multiple- choice tests to the use of alternative forms of assessment. The most recent survey of trends in statewide student assessment programs (Roeber, Bond, and Braskamp, 1997) reveals that 46 of the 50 states have some type of statewide assessment.

C. Purposes and stakes of state assessment programs

The two most popular purposes for assessment according to respondents to a survey of state assessment practices were the "improvement of instruction" -- which was mentioned by 43 states, and "program evaluation" -- mentioned by 38 states. Some other reported purposes (stated in order of number of states mentioning the purpose) include school performance reporting (33), student diagnosis (27), high school exit requirement (17), and school accreditation (11). (Roeber et al., 1997).

However, measurement experts have suggested for some time that "tests used primarily for curriculum advancement will look very different from those used for accountability" (Anderson, 1985, p. 24) and they will have different intended and actual impacts. Likewise, tests used for high stakes decisions (e.g. high school graduation and merit pay) are likely to have different impacts than those used for low stakes decisions (e.g. planning specific classroom interventions for individual students).

It will not always be possible to keep purpose and stakes issues separated when discussing consequences. However, when evidence (or conjecture) about consequences applies to only a specific purpose or level of stakes I will try to make that clear.

D. Format questions

A fairly hot issue in recent years is whether the format of the assessment should vary depending on purposes and whether assessments using different formats have different consequences. In the Roeber et al., (1997) survey of states, multiple choice items were used by 41 states, extended response item types were used by 36, short written response by 24, examples of student work by 10, and what was labeled as performance assessments were used by 9 states. Four states used what were termed projects.

Performance assessment advocates have claimed that the format is important and that positive consequences come from such a format and negative consequences come from multiple-choice formats. Others, like me, are less sure of either of these positions. As with purpose and stakes issues, it will not be possible to always keep these format issues separated in this presentation, but attempts will be made.

Because in recent years more positive claims have been made for "performance assessments" many of the recent attempts to gather consequential evidence have been based on assessments that have used performance assessments. However, I tend to agree with Haney and Madaus when they suggested "...what technology of assessment is used probably makes far less difference than how it is used" (1989, p. 687).

III. EVIDENCE ON THE CONSEQUENCES OF ASSESSMENT PROGRAMS

Lane (1997) developed a comprehensive framework for evaluating the consequences of assessment programs -- concentrating primarily on performance-based assessments. She suggested that both the negative and positive consequences need to be addressed and that one needs to consider both intended and plausible unintended consequences.

For purposes of this presentation, I will discuss some major potential benefits and dangers as follows:

- A. Curricular and instructional reform: Good, bad, or nonexistent?
- B. Motivation/morale/stress/ethical behavior of teachers: Increase of decrease?
- C. Motivation and self-concepts of students: Up or down?
- D. True improvement in student learning, or just higher test scores?
- E. Restore public confidence or provide data for critics?

Evidence on consequences is somewhat sketchy, but the "Lansing State Journal" did report one result in big headlines: "Test Results make School Chief Smile" (Mayes, 1997, p. 1) Many of you have seen such headlines in your own states. When scores go up, the administrators are happy and act as if that means achievement has gone up. It may have, but note well that the consequence I am reporting here (tongue in cheek) is that the superintendent smiled -- not that achievement had improved!

In general there is much more rhetoric than evidence about the consequences of assessment and "too often policy debates emphasize only one side or the other of the testing effects coin" (Madaus, 1991, p. 228). Baker et al., in an article on policy and validity for performance based assessment reported that "less than 5% of the literature cited empirical data." (1993, p. 1213). As they pointed out,

> Most of the arguments in favor of performance-based assessment ... are based on single instances, essentially hand-crafted exercises whose virtues are assumed because they have been developed by teachers or because they are thought to model good instructional practice. (Baker et al., 1993, p. 1211).

I would conclude, as Baker et al. did, that "a better research base is needed to evaluate the degree to which newly developed assessments fulfill expectations" (1993, p. 1216). Koretz suggested that:

> Despite the long history of assessment-based accountability, hard evidence about its effects is surprising sparse, and the little evidence that is available is not encouraging. ...The large positive effects assumed by advocates...are often not substantiated by hard evidence.... (Koretz, 1996, p. 172.).

Reckase (1997) pointed out one of the logical problems in obtaining evidence on the consequences of assessments. The definition of a consequence implies a cause and effect relationship, but most of the evidence has not been gathered in a manner that permits a scholar (or anyone else with common sense) to draw a causative inference.

Green (1997) mentioned many problems in doing research on the consequences of assessment. Among them are that few school systems will welcome reports of unanticipated negative consequences, so cooperation may be hard to obtain; there will be disagreements about the appropriate criterion measures of the consequences; cause-effect conclusions will be disputed; and much of the research undertaken is likely to be undertaken by those trying to prove that what exists is inferior to their new and better idea.

Much of the research has been based on survey information from teachers and principals and, as many authors have pointed out, classroom observations might be more compelling information (see, for example, Linn, 1993 and Pomplun, 1997). Research by McDonnell and Choisser (1997) employed three data sources: face to face interviews, telephone interviews, and assignments collected from the teachers along with a one- page log for each day in a two week period. As the authors pointed out, instructional artifacts is a relatively new strategy --not likely as good as classroom observations, but not as expensive either.

Although evidence is sketchy, there is some! I will discuss such evidence under the headings given earlier regarding the possible dangers and potential benefits of assessment programs.

A. Curricular and instructional reform: Good, bad or nonexistent?

Curricular and instructional reform typically means changing the content of the curriculum or the process of instruction. Not quite fitting either of those categories is changing the length of the school day or the school year. Pipho (1997) has reported that one change between the first and second year of state assessments in Minnesota was the addition of summer school offerings, Saturday classes, and after- school remedial programs. That kind of "reform" is mentioned at other places in the literature, and is, at least arguably, a valuable consequence.

With respect to the more traditional meanings of curricular and instructional reform, it has been commonly assumed that assessments (at least high stakes assessments) will influence curriculum and instruction. One often hears the mantra that "WHAT YOU TEST IS WHAT YOU GET." Taleporos stated flatly at last year's AERA session that "we all know that how you test is how it gets taught." (Taleporos, 1997, p. 1). Actually, the evidence for a test's influence on either curricular content or instructional process is not totally clear. And it will vary by the stakes. Porter, Floden, Freeman, Schmidt, and Schwille reported more than ten years ago that:

> Another myth exposed as being only a half truth is that teachers teach topics that are tested. Little evidence exists to support the supposition that national norm-referenced, standardized tests administered once a year have any important influence on teachers' content decisions. (Porter, et al., 1986, p. 11).

But the "myth" persists. Is it a half truth, a full truth, or just wrong? Logic suggests it may depend on stakes, rigor of the standards, and just what the content is (Airasian, 1988). Some anecdotal evidence also supports the importance of stakes. For example, Floden (personal communication, 1998) states that while in the Content Determinants work (the Porter et al. study just cited) few teachers paid attention to the tests, he is now working in districts where losing accreditation is a real threat, so teachers are busy aligning curriculum to the Michigan Educational Assessment Program (MEAP) and setting aside time before the test for MEAP-specific work.

1. Impact of multiple-choice minimum competency tests on curriculum and instruction

Minimum competency tests were not primarily designed to "reform" the curriculum. Rather, they supposedly measured what schools were already teaching. The tests were intended to find out whether students had learned that material and, if not, to serve as motivators for both the students and the educators. The intended curricular/instructional effect was to concentrate more on the instruction of what was considered to be very important educational goals.

Some earlier writings on the impact of multiple-choice tests suggested that the tests resulted in teachers narrowing the curriculum and corrupting teaching because teachers turned to simply passing out multiple-choice question work sheets. The critics argued that education was harmed due to the narrowing of the curriculum and the teaching and testing for only low level facts.

Aside from the confusion of test format with test content (true measurement experts realize that multiple choice tests are not limited to testing facts), there is insufficient evidence to allow any firm conclusion that such tests have had harmful effects on curriculum and instruction. In fact, there is some evidence to the contrary. Kuhs, Porter, Floden, Freeman, Schmidt, and Schwille (1985) reported that:

> the teachers' topic selection did not seem to be much influenced by the state minimum competencies test or the district-used standardized tests (Kuhs, et al., 1985, p. 151).

There is no evidence of which I am aware showing that fewer high level math courses are taught (or taught to fewer students) in states where students must pass a low level math test in order to receive a high school diploma.

There are a few studies (quoted over and over again) which presumably show that elementary teachers align instruction with the content of basic skills tests (e.g. Madaus, West, Harmon, Lomax and Viator, 1992; Shepard, 1991). And I believe those studies have some validity. It is hard to believe that tests with some stakes connected to them will not have some influence on curriculum and instruction. For example, Smith and Rottenberg (1991) reported on "an extensive research study." "The consequences of external testing were inferred from an analysis of the meanings held by participants and direct observation of testing activities..." (p. 7). They concluded, among other things that (1) external testing reduces the time available for ordinary instruction, (2) testing affects what elementary schools teach -- in high stakes environments, schools neglect material that external tests exclude, (3) external testing encourages use of instructional methods that resemble tests, and (4) "as teachers take more time for test preparation and align instruction more closely with content and format, they diminish the range of instructional goals and activities" (1991, p. 11).

Thus, there are studies suggesting that multiple-choice tests result in a narrowing of the curriculum and more drill work in teaching. But, in fact, the studies are few in number and critics of traditional basic skills testing accept the studies somewhat uncritically. In my opinion, the evidence is not as strong as the rhetoric of those reporting the research would suggest and there is some research evidence that teachers do not choose topics based on the test content (Kuhs et al., 1985).

Green (1997) discussed the evidence and questioned the conclusion that multiple choice tests are harmful stating that "I believe that the data just cited opens to question the assertions about the evils of multiple-choice tests." (Green, 1997, p. 4).

2. Impact of performance assessments on curriculum and instruction

Much of the recent research and rhetoric has been concerned with the effects of performance assessments. Performance assessments are popular in part just because of their supposed positive influence on curricular and instructional reform. Advocates of performance assessment treat as an established fact the position that teaching to traditional standardized tests has "resulted in a distortion of the curriculum for many students, narrowing it to basic, low-level skills" (Herman, Klein, Heath, and Wakai, 1994, p. 1). Further, professional educators have been pushing for curricular reform, suggesting that previous curricula were inadequate and, generally, focused too much on the basics. The new assessments should be more rigorous and schools should be held responsible for these more rigorous standards. As a SouthEastern Regional Vision for Education (SERVE) document entitled "A new framework for state accountability systems" (September 8, 1994) pointed out, some legislative initiatives

> ignored a basic reality: Those schools that had failed to meet older, less rigorous standards were no more able to meet higher standards when the accountability bar was raised. As a result, state after state is confronted with previously failing schools failing the new systems (SERVE, 1994, p. 2).

What does the research tell us about the curricular and instructional effects of performance assessments? Khattri, Kane, and Reeve visited sixteen schools across the United States that were developing and implementing performance assessments. They interviewed school personnel, students, parents and school board members; collected student work; and conducted observations. They concluded that:

> In general, our findings show that the effect of assessments on the *curriculum* teachers use in their classrooms has been marginal, although the impact on *instruction* and on *teacher roles* in some cases has been substantial (Khattri, Kane, and Reeve, 1995, p. 80).

Chudowsky and Behuniak (1997) used teacher focus groups from seven schools representing a cross section of schools in Connecticut. These focus groups discussed their perceptions of the impact of the Connecticut Academic Performance Test -- an assessment that uses multiple-choice, grid-in, short answer and extended response items. Teachers in all but one of the schools reported that preparing students and aligning their instruction to the test "resulted in a narrowing of the curriculum" (Chudowsky and Behuniak, 1997, p. 8). Regarding instructional changes, "teachers most frequently reported having students 'practice' for the test on CAPT sample items" (p. 6). However the schools also reported using strategies "to move beyond direct test preparation into instructional approaches" (p. 6). Teachers also

> consistently reported that the most negative impact of the test is that it detracts significantly from instructional time. Teachers at all of the schools complained vehemently about the amount of instructional time lost to administer the test (p. 7).

Koretz, Mitchell, Barron, and Keith (1996) surveyed teachers and principals in two of the three grades in which the Maryland School Performance Assessment Program (MSPAP) is administered. As they reported, the MSPAP program "is designed to induce fundamental changes in instruction" (p. vii). While about three-fourths of the principals and half of the teachers expressed general support for MSPAP, fifteen percent of the principals and 35% of the teachers expressed opposition. One interesting finding was that about 40% of fifth-grade teachers "strongly agreed that MSPAP includes developmentally inappropriate tasks" (p.

viii). One of the summary statements made by Koretz Mitchell, Barron, and Keith (1996) is as follows:

> The results reported here suggest that the program has met one of its goals in increasing the amount of writing students do in school. At the same time, teachers' responses suggest the possibility that this change may have negative ramifications as well, in terms of both instructional impact and test validity. Many teachers maintain that the emphasis on writing is excessive and that instruction has suffered because of the amount of time required for writing. ...[also,] emphasis on writing makes it difficult to judge math competence of some students" (Koretz, Mitchell, Barron, and Keith, 1996, p. xiii).

Rafferty (1993) surveyed urban teachers and staff regarding the MSPAP program. Individuals were asked to respond in Likert fashion to several statements. When the question was "MSPAP will have little effect on classroom practices" 33% agreed or strongly agreed, 24% were uncertain, 42% disagreed or strongly disagreed and 1% did not respond. To the statement "classroom practices are better because of MSPAP" 21% were in agreement, 36% were uncertain, 41% disagreed, and 2% did not answer. To the statement "MSPAP is essentially worthwhile," 24% agreed or strongly agreed, 25% were uncertain, and 48% disagreed or strongly disagreed (3% did not respond). Perhaps a reasonable interpretation of these data is that MSPAP will likely have an impact, but not necessarily a good one.

Koretz, Barron, Mitchell, and Stecher (1996) did a study for Kentucky similar to the one done by Koretz, Mitchell, Barron, and Keith in Maryland. They surveyed the teachers and principals in Kentucky regarding the Kentucky Instructional Results Information System (KIRIS) and found much the same thing as had been found in Maryland. Among other findings, were the following (Note 3) :

> --90% of the teachers agreed that portfolios made it difficult to cover the regular curriculum (p. 37); --most teachers agreed that imposing rewards and sanctions causes teachers to ignore important aspects of the curriculum (p. 42); --portfolios were cited as having negative effects on instruction almost as often as having had positive effects (p. xi); --almost 90% of the teachers agreed that KIRIS caused them to de- emphasize or neglect untested material (p. xiii); and --other aspects of instruction have suffered as a result of time spent on writing, and emphasis on writing makes it difficult to judge the mathematical competence of some students (p. xv).

McDonnell and Choisser (1997) studied the local implementation of new state assessments in Kentucky and North Carolina. They concluded that

> Instruction by teachers in the study sample is reasonably consistent with the state assessment goals at the level of classroom activities, but not in terms of the conceptual understandings the assessments are measuring. Teachers have added new instructional strategies ... but ... they have not fundamentally changed the depth and sophistication of the content they are teaching. (1997, p. ix.).

Stretcher and Mitchell (1996) reported on the effects of portfolio-driven reform in Vermont and stated that

> The Vermont portfolio assessment program has had substantial positive effects on fourth-grade teachers' perceptions and practices in mathematics. Vermont teachers report that the program has taught them a great deal about mathematical problem solving and that they have changed their instructional practices in important ways (Stretcher and Mitchell, 1996, p. ix).

Smith, Nobel, Heinecke, Seck, Parish, Cabay, Junker, Haag, Tayler, Safran, Penley, and Bradshaw (1997) conducted a study of the consequences of the, now discarded, Arizona Student Assessment Program (ASAP). Although the program had several parts including some norm referenced testing with the Iowa tests, the most visible portion of ASAP was the performance assessment. Teacher opinion of the direction of the effect of ASAP on the curriculum was divided.

> Some defined 'ASAP' as representing an unfortunate and even dangerous de-emphasis of foundational skills, whereas others welcomed the change or saw the new emphasis as encompassing both skills and problem solving. (Smith et al., 1997, p. 40).

Some interesting quotes by teachers found in the Smith et al. report are as follows:

> Nobody cares about basics...The young teachers coming out of college will just perpetuate the problem since they are learning whole language instruction and student-centered classroom. Certainly these concepts have their merits, but not at the expense of basics on which education is based (p. 41).
>
> The ASAP ... is designed to do away with 'skills' because kids today don't relate to skills, because they are boring. By pandering to this we are weakening our society, not strengthening it. *It is wrong*! I was told by a state official that teachers would be more like coaches under ASAP. Ask any coach if they teach skills in isolation before they integrate it into their game plan. They will all tell you yes. I rest my case. (Smith et al., 1997, p. 41).

BEST COPY AVAILABLE

As Smith, et al. report, about two thirds of the teachers believed that "pupils at this school need to master basic skills before they can progress to higher order thinking and problem solving" (1997, p. 41). Forty three percent of the teachers believed that "ASAP takes away from instructional time we should be spending on something more important." (p. 44). In spite of many teachers being unhappy with the content of ASAP, "about 40% of the teachers reported that district scope and sequences had been aligned with ASAP." (p. 46). As the authors report, "changes consequent to ASAP seemed to fall into a typology that we characterized as 'coherent action,' 'compliance only,' 'compromise,' and 'drag.'" (p. 46).

Miller (1998) studied the effect of state mandated performance based assessments on teachers' attitudes and practices in six different contexts (grade level and content areas). Two questions were asked relevant to curricular and instructional impacts.

> "I have made specific efforts to align instruction with the state assessments." (Percents who agreed or strongly agreed ranged from 54.5 to 92.7% across the six contexts.) "I feel that state mandatory assessments have had a negative impact by excessively narrowing the curriculum covered in the classroom." (Percents who agreed or strongly agreed ranged from 28.7 to 46.8%. Only teachers in five of the contexts responded to that question.)

The two questions provide interesting results. While the majority made specific efforts to align instruction, the majority did not feel it resulted in excessively narrowing of the curriculum. However, as Miller pointed out, "the assessments were usually intended to give supplemental information. Consequently, they do not reflect everything that students learn, and only provide a small view of student performance..." (Miller, 1998, pp. 5-6). To align instruction to assessments that provide only a small view of student performance without excessively narrowing the curriculum would seem to be a difficult balancing act.

While more research and opinions could be reviewed, a reasonably summary is that if stakes are high enough and if content is deemed appropriate enough by teachers, there is likely to be a shift in the curriculum and instruction to the content sampled by the test (or the content on the test if the test is not secure). If stakes are low, and/or if teachers believe the assessment is testing developmentally inappropriate materials and/or teaching to the assessment would reduce the amount of time the teachers wish to spend on other -- what they consider more important -- content, the impact is not so obvious.

B. Motivation/morale/stress/ethical behavior of teachers:

Increase or Decrease?

Many would argue, quite reasonably, that if we are to improve education, we must depend on the front line educators -- the teachers -- to lead the charge. Do large scale assessments tend to improve the efforts, attitudes and ethical behavior of teachers?

Smith and Rottenberg suggested that external tests negatively affect teachers. As they wrote:

> the chagrin they felt comes from their well-justified belief that audiences external to the school lack interpretive context and attribute low scores to lazy teachers and weak programs (Smith and Rottenberg, 1991, p. 10).

Although I believe they were primarily discussing the effects of traditional assessments, one should expect the same reaction from performance assessments. Audiences external to the school are no more able to infer correct causes of low scores on performance assessments than they are to infer correct causes of low scores on multiple-choice assessments. The inference to lazy teachers and weak programs is equally likely no matter what the test format or test content.

Koretz, Mitchell, Barron and Keith (1996) reported that for the Maryland School Performance Assessment Program (MSPAP):

> Few teachers reported that morale is high, and a majority reported that MSPAP has harmed it. ... 57% of teachers responded that MSPAP has led to a decrease in teacher morale in their school, while only a few (4%) reported that MSPAP has produced an increase (Koretz, Mitchell, Barron and Keith, 1996, p. 24).

Koretz, Barron, Mitchell and Stecher (1996) in the Kentucky study reported that "about 3/4 of teachers reported that teachers' morale has declined as a result of KIRIS" (p. x). Stecklow (1997) reported that there were conflicts in over 40% of Kentucky schools about how to divide up the reward money. So affect was not necessarily high even in the schools that got the rewards! Koretz, Barron, Mitchell and Stecher (1996) also found that principals reported that KIRIS had affected attrition. But the attrition was for both good and poor teachers.

With respect to effort, at least the teachers in Kentucky reported that their efforts to improve instruction and learning had increased (Koretz, Barron, Mitchell and Stecher, 1996, p. 23). But at some point, increased efforts lead to burnout -- and thus attrition increases.

It is commonly believed that some teachers engage in behaviors of questionable ethics when teaching toward, administering and scoring high stakes multiple-choice tests. What about performance assessments? Koretz, Barron, Mitchell and Stecher (1996) reported that in Kentucky

> Appreciable minorities of teachers reported questionable test-administration practices in their schools. About one-third reported that questions are at least occasionally rephrased during testing time, and

> roughly one in five reported that questions about content are answered during testing, that revisions are recommended during or after testing, or that hints are provided on correct answers. (Koretz, Barron, Mitchell and Stecher, 1996, p. xiii).

In summary, the evidence regarding the effects of large scale assessments on teacher motivation, morale, stress and ethical behavior is sketchy. But what evidence there is, coupled with what seems logical, suggests that increasing the stakes for teachers will increase efforts, lead to more burnout, decrease morale, and increase the probability of unethical behavior.

C. Motivation and self-concepts of students: Up or down?

With respect to assessment impacts on the affect of students, we are again in a subarea where there is not a great deal of empirical evidence. Logic suggests that the impact on students may be quite different for those tests where the stakes apply to them than for tests where the stakes impact the teachers. Impact surely depends on whose ox is getting gored by the stakes.

Also, the impact should depend on how high the standards are. It is reasonably to believe that the impact of minimal competency tests would be minimal for the large majority of students for whom such tests would not present a challenge. However, for those students who had trouble getting over such a minimal hurdle, the tests probably would both increase motivation and increase frustration and stress -- the exact mix varying on the personality characteristics of the students.

Smith and Rottenberg found that for younger students teachers *believed* that standardized tests

> cause stress, frustration, burnout, fatigue, physical illness, misbehavior and fighting, and psychological distress (Smith and Rottenberg, 1991, p. 10).

That belief of teachers may be true, but certainly does not constitute hard evidence. I come closer to Ebel's view when he suggested that

> Of the many challenges to a child's peace of mind caused by such things as angry parents, playground bullies, bad dogs, shots from the doctor, and things that go bump in the night, standardized tests must surely be among the least fearsome for most children (Ebel, 1976, p. 5).

Lane and Parke (1996), reporting on the consequences of a math performance assessment found that some students developed feelings of inadequacy and, as a result, were less motivated. Miller (1998) found that the percent of teachers responding positively to the statement that performance assessments "increased student confidence" ranged from only 9.1 to 37.6% across five different contexts.

However, Kane et al. (1997) employing a qualitative, case-study methodology and visiting 16 schools ("not confirmed to be representative" --p. xvi) developing and implementing performance assessments reported that

> many interviewees reported that students exhibit a greater motivation to learn and a greater amount of engagement with performance tasks and portfolio assignments than with other types of assignments" (Kane et al., 1997, p. 201).

Koretz, Barron, Mitchell and Stecher (1996) reported that in Kentucky one third of the teachers reported that students' morale had deteriorated and virtually none reported an increase in student morale. They also reported that an emphasis or writing caused students to become tired of writing.

As mentioned earlier, one of the factors effecting student affect is how high the standards are set. Minimum standards are not likely to have a major impact. High standards might. Linn (1994) has pointed out that

> The dual goals of setting performance standards for student certification that are both 'world class' and apply to 'all' students are laudable, but it cannot simply be assumed that only positive effects will result from this press (Linn, 1994, p. 8).

Linn quoted Coffman as follows:

> Holding common standards for all pupils can only encourage a narrowing of educational experiences for most pupils, doom many to failure, and limit the development of many worthy talents (Coffman, 1993, p. 8; quoted from Linn, 1994).

We are simply putting too many students and too many teachers under too much pressure if we hold unrealistically high standards for all students. As Bracey has said, in an article entitled "Variance Happens -- get over it!"

> We are currently in a period that adheres rabidly to an all-children-can-learn philosophy. ... The stance is a philosophical, moral -- almost religious -- posture taken by a wide spectrum of educators and psychologists who ought to know better. ... By telling everyone that all children can learn, we set the stage for the next great round of educational failure when it is revealed that not everyone **has** learned, in spite of our sincere beliefs and improved practices. (Bracey, 1995, p. 22 and 26).

Of course his point is not that some children can not learn anything, but that not everyone can achieve at high standards in academics anymore than everyone can become athletically proficient enough in every sport to play on the varsity teams.

D. True improvement in student learning, or just higher test scores?

> *In mandating tests, policy makers have created the illusion that test performance is synonymous with the quality of education (Madaus, 1985, p. 617).*

BEST COPY AVAILABLE

All of us recognize that it is possible for test scores to go up without an increase in student learning on the domain the test supposedly samples. This occurs, for example, when teachers teach the questions on non-secure tests. Teaching too closely to the assessment results in the inferences from the test scores being corrupted. One can no longer make inferences from the test to the domain. The Lake Wobegon effect results. Many of us have written about that (e.g. Mehrens and Kaminski, 1989).

If the assessment questions are secure and the domain the test samples is made public, corrupting reasonable inferences from the scores is more difficult. If the inference from rising test scores of secure tests is that students have learned more of the domain the test samples, that is likely a correct inference. However, those making inferences may not realize how narrow the domain is, or that a test sampling a similar sounding but somewhat differently defined domain might give different results. Of course, if the inference from rising scores is that educational quality has gone up, that may not be true.

1. Improvement on traditional tests

Pipho has reported that:

> Ironically, every state that has initiated a high school graduation test in grade 8 or 9 has reported an initial failure rate of approximately 30%. By 12th grade, using remediation and sometimes twice-a-year retests, this failure rate always gets down to well under 5% (Pipho, 1997, p. 673).

Is this *true improvement*, or is it a result of teaching to the test? Recall that these tests are supposedly secure so one cannot teach the specific questions. However, one could limit instruction to the general domain the tests sample. My interpretation is that the increase in scores represents a true improvement on the domain the test samples, but that it does not necessarily follow that it is a true improvement in the students' education.

2. Improvement on performance assessments

What about performance assessments? Do increases in scores indicate necessary improvement in the domain, or an increase in educational quality? Certainly no more so than for multiple-choice assessments, and perhaps less so. Even if specific tasks are "secure," performance assessments are generally thought to be even more "memorable" and reusing such assessments can result in corrupted inferences. If the inference is to only the specific task, there may not be too much corruption, but any inferences to a domain the task represents or to the general quality of education are as likely to be incorrect for performance assessments as for multiple-choice assessments.

Shepard, Flexer, Hiebert, Marion, Mayfield, and Weston (1996) conducted a study investigating the effects of classroom performance assessments on student learning. As they stated:

> Overall, the predominant finding is one of no-difference or no gains in student learning following from the year-long effort to introduce classroom performance assessments. Although we argue subsequently that the small year-to-year gain in mathematics is real and interpretable *based on qualitative analysis, honest discussion* of project effects must acknowledge that *any benefits are small and ephemeral* (Shepard et al., 1996, p. 12, emphasis added).

Others, doing less rigorously controlled studies based on teacher opinion surveys, have been equally cautious in their statements. Khattri et al. (1995), in their study visiting 16 schools stated that

> Only a few teachers said performance-based teaching and assessment helped students learn more and develop a fuller multi-disciplinary understanding (Khattri et al., 1995, p. 82).

Koretz, Barron, Mitchell and Stecher (1996) reported that

> Few teachers expressed confidence that their own schools' increases on KIRIS were largely the results of improved learning (Koretz, Barron, Mitchell and Stecher, 1996, p. xiii)

The authors go on to suggest that

> A variety of the findings reported here point to the possibility of inflated gains on KIRIS--that is, the possibility that scores have increased substantially more than mastery of the domains that the assessment is intended to represent (Koretz, Barron, Mitchell and Stecher, 1996, p. xv).

Kane et al. (1997) concluded from their study that

> In the final analysis, the success of assessment reform as a tool to enhance student achievement remains to be rigorously demonstrated (Kane et al., 1997, p. 217).

Miller (1998) asked teachers whether they believed the state mandated performance assessments "have had a positive effect on student learning." Percents across five contexts ranged from 11.3% to 54.7%. When asked whether the tests results were "an accurate reflection of student performance" the percentages ranged from 13.1% to 28.7%.

Finally, for some types of portfolio assessments, one does not even know who did the work. As Gearhart, Herman, Baker, and Whittaker pointed out: "This study raises questions concerning validity of inferences about student competence based on portfolio work." (Gearhart et al., 1993, p. 1).

3. Conclusions about increases

In conclusion, there is considerable evidence that students' pass rates increase on secure high-stakes (mostly multiple-choice) graduation tests. There is at least some reason to believe that students have increased their achievement levels on the specific domains the secure tests are measuring. (Of course, if supposedly secure tests are not actually secure the inference that increased scores indicate increased achievement could be incorrect.) There is less evidence about increases in scores for performance assessments. While it is true that some states (e.g. Kentucky) have shown remarkable gains in scores, evidence points to the possibility that the gains are inflated and there is generally less confidence that achievement in the represented domain has also increased. In neither case can we necessarily infer that quality of education has increased. That inference cannot flow directly from the data. Rather, it must be based on a philosophy of education that says an increase in the domain tested represents an increase in the quality of education. As Madaus stated, it is an illusion to believe at an abstract level that test performance is *synonymous with* quality of education. Nevertheless, test performance can *inform us about* the quality of education -- at least about the quality of education on the domain being assessed.

E. Restore public confidence or provide data for critics?

At an abstract level, it seems philosophically wrong and politically shortsighted for educators to argue against the gathering of student achievement data for accounting and accountability purposes. My own belief is that an earlier stance of the NEA against standardized tests resulted in the public wondering just what it was the educators were trying to hide. I suspect the NEA stance contributed to the action in the state departments that Womer mentioned in 1984. Certainly the public has a right to know something about the quality of the schools they pay for and the level of achievement their children are reaching in those schools.

Some educators strongly believe -- with some supporting evidence -- that the press has incorrectly maligned the public schools (e.g. Bracey, 1996, and Berliner and Biddle, 1995). While their views have not gone unchallenged (see Stedman, 1996) it does seem true that bad news about education travels faster than good news about education. Will the data from large scale assessments change the public's views?

The answer to the above question depends, in part, upon whether the scores go up, go down, or stay the same. It also depends upon whether the public thinks we are measuring anything worthwhile. And it also depends upon how successful we are at communicating the data and communicating what reasonable inferences can be drawn from the data.

Scores generally have been going up in Kentucky, but it has not resulted in all the press highlighting the great job educators are doing in Kentucky. For example Stecklow (1997), in writing about the Kentucky approach suggests:

> It has spawned lawsuits, infighting between teachers and staff, anger among parents, widespread grade inflation -- and numerous instances of cheating by teachers to boost student scores. (Stecklow, 1997, p. 1).

A conclusion of the evaluation done of KIRIS by The Evaluation Center of Western Michigan University stated that

> ...**all** the cited evidence suggests stakeholders have questions concerning the legitimacy, validity, reliability, and fairness of the KIRIS assessment. We have no evidence to suggest that parents think the assessment component of KIRIS is a fair, reliable, and valid system (The Evaluation Center, 1995, p. 20).

The Stecklow quote is not a ringing endorsement of the program or the quality of education in Kentucky. The Evaluation Center quote is not a ringing endorsement of the quality of the data in the assessment. But, in general, the public is happier with high scores than they are with low scores -- often considering which district to live in based on published test scores. (The public may be making two quite different inferences from these scores. One, probably an incorrect inference, is that the district with the higher scores has better teachers or a better curriculum. A second, correct inference is that if their children attend the district with the higher scores they will be more likely to be in classes with a higher proportion of academically able fellow students.)

There is currently considerable concern about whether the newer "reform" assessments cover the correct content. Reform educators were not happy with minimum competency tests covering basics; but the public is not happy with what they perceive to be a departure from teaching and testing the basics. Baker, Linn, and Herman (1996) talk about the crisis of credibility that performance assessments suffer based on a large gap between the views of educational reformers and segments of the public. McDonnell (1997) stated that

> ...the political dimensions of assessment policy are typically overlooked. Yet because of their link to state curriculum standards, these assessments often embody unresolved value conflicts about what content should be taught and tested, and who should define that content. (McDonnell, 1997, p. v).

As McDonnell pointed out, there are fundamental differences between what educational reformers and large segments of the public believe should be in the curriculum.

> The available opinion data strongly suggest that the larger public is skeptical of new curricular approaches in reading, writing, and

mathematics (McDonnell, 1997, p. 67).

The truth of this can be seen by the fight over the mathematics standards in California.

The press and public seem either reasonably unimpressed by the data educators provide and/or make incorrect inferences from it. What can change that? We need to gather high quality data over important content and communicate the data to the public in ways that encourage correct inferences about students' levels of achievement. We need to be especially careful to discourage the public from making causative inferences if they are not supported by the research data.

IV. ARE THE CONSEQUENCES GOOD OR BAD?

It should be obvious by now that I do not believe we have a sufficient quantity of research on the consequences of assessment. Further, the evidence we do have is certainly not of the type from which we can draw causative inferences, which seems to be what the public wants to do. Given the evidence we do have, can we decide if it suggests the consequences of assessment are positive or negative? If the evidence were better, could we decide if the consequences are positive or negative? I maintain that each of us can decide, but we may well disagree. Interpreting the consequences as being good or bad is related to differences in convictions about the proper goals of education. Let us look at the evidence regarding each of the five potential benefits and/or dangers with respect to the quality of the consequences.

A. Curricular and Instructional Reform

While there is no proven cause and effect relationship between assessment and the curriculum content or instructional strategies there is some evidence and compelling logic to suggest that high stakes assessments can influence both curriculum and instruction. Is this good or bad? It is a matter of one's goals. Reform educators were dismayed to think that minimum competency tests using multiple-choice questions were influencing curriculum and instruction. They pushed for performance assessments, not because they abhorred tests influencing curriculum and instruction, but because they wanted the tests to have a different influence.

The public was not dismayed that educators tested the basics -- they rather approved. They believe (some evidence suggests incorrectly) that educators have moved away from basics and are dismayed. Obviously the narrowing and refocusing of the curriculum and instructional strategies are viewed as either negative or positive depending upon whether the narrowing and refocusing are perceived to be toward important content. Educators and the public do not necessarily agree about this.

B. Increasing Teacher Motivation and Stress

Increasing teacher stress may be perceived as good or bad -- depending on whether one believes teachers are lazy and need to be slapped into shape or whether one believes (as I do) that teachers already suffer from too much job stress.

C. Changing Students' Motivation or Self Concept

We might all favor an increase in student motivation. I, for one, do not believe a major problem in education in the United States is that students are trying too hard to learn too much. But some educators do worry about the stress that tests cause students (recall the quote from Smith and Rottenberg). There is such a thing as "test anxiety" (more accurately called evaluation anxiety), but many would argue that occasional state anxiety is a useful experience -- perhaps helping individuals to learn how to cope with anxiety and to treat stress as eustress rather than distress.

But what if assessment lowers students' self concepts? Again, this could be either good or bad -- depending on whether one believes students should have a realistic view of how inadequate their knowledge and skills are. (Recall that in Japan, whose students outperform U.S. students, the students do not feel as confident in their math competencies as do U.S. students.) As one colleague has pointed out to me, we are not necessarily doing students a favor by allowing them to perceive themselves as competent in a subject matter if that, indeed, is not the truth (Ryan, personal communication, 1997).

D. Increased Scores on Assessments

Surely this is good -- right? It again depends. It depends on whether the gains reflect improvement on the total domain being assessed or just increases in scores, whether we care about the tested domain, and whether, as a result of the more focused instruction, other important domains (not being tested) suffer.

E. Public Awareness of Student Achievement

Is public awareness of how students score on assessments good or bad? Obviously one answer is that it depends on whether valid inferences are drawn from the data. One part of the validity issue is whether the scores truly represent what students know and can do. Another part of the validity issue is whether the public draws causative inferences that are not supported by the data.

In addition to the question of whether the inferences are valid, there is the issue of how the public responds. Would negative news stimulate increased efforts by the public to assist educators --e.g. by trying to ensure children start school ready to learn, by providing better facilities, by insisting their children respect the teachers? Or would negative news result in more rhetoric regarding how bad public schools are, how bad the teachers are, and how we should give up on them and increase funding to private schools at the expense of funding public schools? Would positive news result in teachers receiving public accolades and more respect or would the public then place public education on a back burner -- because the "crisis" was over?

While I come down on the side of giving the public data about student achievements, the communication with the public must be done with great care. I believe there is a propensity for the public (at least the press) to engage in inappropriate blaming of educators when student achievement is not as high as desired. I am reminded of Browder's (1971) suggestion that accountability boils down to who gets hanged when things go wrong and who does the hanging. Educators have good reason to believe that they are the ones who will get hanged and the public, abetted by the press, will do the hanging.

Dorn has stated that "test results have become the dominant way states, politicians, and newspapers describe the performance of schools." (1998, p. 2). He was not suggesting that was a positive happening.

V. WHAT VARIABLES CHANGE PROBABILITIES FOR GOOD OR BAD IMPACTS

Since *whether* consequences are good or bad is partly a matter of one's educational values, it is difficult to answer this question. Nevertheless, I will provide a few comments.

A. Impact should be (and likely is) related to purposes.

As has been mentioned, there are two major purposes of large scale assessments: to drive reform, and to see if reform practices have had an impact on student learning. These are somewhat contradictory purposes, because current reformers believe assessment should be "authentic" if it is to drive reform and most authentic assessment is not very good measurement -- at least by any conventional measurement criteria.

B. Impact (and purposes) are likely related to test content, and the *public involvement* in determining content and content standards.

> Successful assessment reform needs to be an open and inclusive process, supported by a broad range of policy makers, educators, and the public, and closely tied to standards in which parents and the community have confidence. (The CRESST Line, 1997, p. 6).

The impact is not likely to be positive in any overall sense if the public has not bought into the content standards that are being assessed.

One can also expect some problems if the content and test standards are set too high. The politically correct rhetoric that "all children can learn to high levels" has yet to be demonstrated as correct. Recall the quote by Coffman given earlier. Recall also Bracey's article entitled: Variance happens--get over it. Or, as a colleague once said, would we require the PE instructor to get all students up to a level where they are playing on the varsity team?

C. Impact may be related to item or test format.

If the issue is whether the overall impact is good or bad, there is not much evidence that item or test format matters. The abstract notion that teaching to improve performance assessment results means educators will be teaching like they should be teaching whereas teaching to improve multiple-choice test scores means teaching is of poor quality is just nonsense.

D. Impact may be related to the quality of the assessment (perceived or real) and the assessment procedures (e.g. test security and reporting practices).

If educators do not believe the assessments provide high quality data, they may not pay much attention to them. Cunningham made this point very forcefully in discussing the Kentucky Instructional Results Information System (KIRIS) program:

> As teachers begin to realize that the test has no legitimacy and that it is too technically deficient to be influenced by how they teach, they will stop paying attention to it. ... Measurement driven instruction does not work when teachers fail to see the connection between measurement and instruction. (Cunningham, no date, p. 2).

Whatever one believes about the technical adequacy of KIRIS, Cunningham's general point would seem accurate: If teachers do not see any connection between the assessment results and their instructional approaches the measurement is unlikely to impact instruction.

Another example comes from a paper Smith et al. (1997) wrote on the consequences of the Arizona Student Assessment Program (ASAP). Some teachers believed that the ASAP skills were not developmentally appropriate (p. 38), some objected to what they perceived as poor-quality rubrics and to the subjectivity of the scoring process (p. 39), some thought ASAP was just a fad and one teacher referred to ASAP as Another Stupid Aggravating Program (p. 43). Again, the point is not whether the teachers' perceptions were correct. But if they perceive the assessment quality to be poor, they are not likely to be very positively impacted by it.

E. Impact may depend upon degree of sanctions.

Some limited research evidence regarding this variable comes from a McDonnell and Choisser (1997) study. They investigated the local implementation of state assessments in North Carolina and Kentucky. As they suggested, Kentucky's program involved high stakes for schools and educators, with major consequences attached to the test results. The North Carolina assessment had no tangible consequences attached to it. However

> teachers in the two state samples perceive the new assessments in much the same way and take them equally seriously. With few exceptions, their teaching reflects the assessment policy goals of their respective states to a similar degree. (McDonnell and Choisser, 1997, p. ix).

Of course, the North Carolina assessments have some consequences. Results are presented in district 'report cards' and in school building improvement results. And, at the time the study was conducted, McDonnell and Choisser reported that

> probably the most potent leverage the assessment system has over the behavior of teachers is the widespread perception that local newspapers plan to report test scores not just by individual school, which has been done traditionally, but also by specific grade level and even by classroom (1997, p. 16).

One can imagine why teachers in North Carolina might think the stakes were fairly high in spite of no state financial rewards or sanctions.

Thus, in spite of the evidence which shows no distinctions between Kentucky, which used financial rewards, and North Carolina, which simply made the scores available, the perceived stakes to the teachers may not have been much different. I continue to believe that as stakes increase, dissatisfaction increases, fear increases, cheating increases, and lawsuits increase. However, efforts may also increase to improve scores and, if the procedures are set up to make it difficult to improve scores without improving competence on the domain, student learning should increase also.

F. Impact may relate to level of professional development.

Unfortunately, many current reform policies concentrate more on standards and assessments than they do the professional development of teachers. In the Arizona SAP program for example, only 19% of the teacher felt that adequate professional development had been provided (Smith et al., 1997). Combs, in his critique of top-down reform mandates stated that: "Things don't change people; people change things." (quoted from Smith et al., 1997, p.50). As Smith et al. pointed out in their review, Cohen (1995, p. 13) had noted the apparent anomaly in the systemic reform movement and accountability intentions. Motivated by perceptions that public schools are failing,

> advocates of systemic reform propose to radically change instruction, and for that they must rely on teachers and administrators. But these agents of change are the very professionals whose work reformers find so inadequate. (Quoted from Smith et al., 1997, p. 105).

VI. CONCLUSIONS

So, what can we conclude about the consequences of assessment? I list a dozen.

A. Purposes and Expectations

1. There are a variety of purposes for and expectations regarding the consequences of assessment. Some of these may be unrealistic. "Evaluation and testing have become *the* engine for implementing educational policy" (Petrie, 1987, p. 175).

B. Need for Evidence

2. Scholars seem to agree that it is unwise, illogical, and unscholarly to just assume that assessments will have positive consequences. There is the potential for both positive consequences and negative consequences.

C. Quantity and Quality of the Evidence

3. It would profit us to have more research.

4. The evidence we do have is inadequate with respect to drawing any cause/effect conclusions about the consequences. If instruction changes concomitant with changes in both state curricular guidelines and state assessments, how much of the change was due to which variable?

D. Evaluating the Evidence

5. Not everyone will view changes (e.g. reforming curriculum in a particular way) with the same affect. Some will think the changes represent positive consequences and others will think the changes constitute negative consequences.

E. Curricular and Instructional Consequences

6. High stakes assessments probably do impact both curriculum and instruction, but assessments alone are not likely as effective as they would be if there were more teacher professional development.

7. Attempts to reform curriculum in ways neither front line teachers nor the public support seems unwise.

F. Impact on Teachers

8. High stakes assessments increase teacher stress and lower teacher morale. This seems unfortunate to me, but may make others happy.

9. Assessments can assist both students and teachers in evaluating whether the students are achieving at a sufficiently high level. This seems like useful knowledge.

G. Impact on Test Scores and Student Learning

10. High stakes assessments will result in higher test scores. Both test security and the opportunity to misadminister or mis-score tests must be considered in evaluating whether higher scores represent increased knowledge. If the test items are secure (and reused items are not memorable), and if tests are administered and scored correctly, it seems reasonably to infer that higher scores indicate increased achievement in the particular domain the assessment covers. That is good if the domain represents important content and if teaching to that domain does not result in ignoring other equally important domains. If tests are not secure, or are incorrectly administered or scored, there is no reason to believe that higher scores represent increased learning.

H. Impact on Public

11. The public and the press are more likely to use what they believe to be "inadequate" assessment results to blame educators than to use "good" results to praise them. They will continue to make inappropriate causative inferences from the data. The public will not be impressed by assessments over reform curricula they consider irrelevant.

I. Confounding Format, Content and Stakes in Considering Consequences

12. There has been a great deal of confounding of item format, test content and the stakes. Which format is used probably makes far less difference than how it is used.

References

Airasian, P. W. (1988). Measurement driven instruction: A closer look. *Educational Measurement: Issues and Practice*, 7(4), 6-11.

Anderson, B. L. (1985). State testing and the educational community: Friends or foes? *Educational Measurement: Issues and Practice*, 4(2), 22-25.

Anderson, B., and Pipho, C. (1984). State-mandated testing and the fate of local control. *Phi Delta Kappan*, 66, 209-212.

Baker, E. L., Linn, R. L., and Herman, J. L. (1996, Summer). CRESST: A continuing mission to improve educational assessment. *Evaluation Comment*.

Baker, E.L., O'Neil, H.F., and Linn, R.L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist*, 48(12), 1210-1218.

Berliner, D., and Biddle, B. (1995). *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. New York: Addison-Wesley.

Bracey, G.W. (Fall, 1995). Variance happens -- get over it! *Technos*, 4(3), 22-29.

Bracey, G.W. (1996). International comparisons and the condition of American education, *Educational Researcher*, 25(1), 5-11.

Browder, L.H. Jr. (1971). *Emerging patterns of administrative accountability*. Berkeley, CA: McCutchan.

Chudowsky, N., and Behuniak, P. (1997, March). Establishing consequential validity for large-scale performance assessments. Paper presented at annual meeting of the National Council of Measurement in Education, Chicago, IL.

CRESST. (1997, Spring). Analyzing statewide assessment reforms. *The CRESST Line*.

Cunningham, G. K. (No date). Response to the response to the OEA panel report. University of Louisville.

Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives*, 6, No. 1 (Entire Issue). (Available online at http://epaa.asu.edu/epaa/v6n1.html).

Ebel, R.L. (1976). The paradox of educational testing. *Measurement in Education*, 7(4), 1-12.

The Evaluation Center. (1995). *An independent evaluation of the Kentucky Instructional Results Information System (KIRIS)* [Report conducted for The Kentucky Institute for Education Research]. Western Michigan University.

Floden, R.E. (1998). Personal communication.

Froomkin, D. (Sept. 29, 1997). National education tests: An introduction. *Back to the top* [on line], Digital Ink Company.

Gearhart, M., Herman, J.L., Baker, E.L., and Whittaker, A.K. (July 1993). *Whose work is it? A question for the validity of large-scale portfolio assessments.* CSE Technical Report 363. Center for the study of evaluation, National Center for Research on Evaluation Standards, and Student Teaching, Graduate School of Education, University of California, Los Angeles.

Green, D. R. (1997, March). *Consequential aspects of achievement tests: A publisher's point of view.* Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, IL.

Haney, W., and Madaus, G. (1989). Searching for alternatives to standardized tests: Whys, whats, and whithers. *Phi Delta Kappan*, 70(9), 683-687.

Herman, J.L., Klein, D.C.D., Heath, T.M. and Wakai, S.T. (December, 1994). *A first look: Are claims for Alternative assessment holding up?* CSE Technical Report 391. National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Jones, L.V. (1997). *National tests and education reform: Are they compatible?* William H. Angoff Memorial Lecture Series. Educational Testing Service.

Kane, M. B., Khattri, N., Reeve, A. L., and Adamson, R. J. (1997). *Assessment of student performance.* Washington D.C.: Studies of Education Reform, Office of Educational Research and Improvement, U.S. Department of Education.

Khattri, N., Kane, M. B., and Reeve, A. L. (1995). *How performance assessments affect teaching and learning* [Research Report]. Educational Leadership.

Koretz, D. (1996). Using student assessments for educational accountability. In E. A. Hanushek and D.W. Jorgenson (Eds). *Improving America's schools: The role of incentives.* (pp. 171-195). National Academy Press, Washington, DC.

Koretz, D, Barron, S., Mitchell, K., and Stecher, B. (1996, May). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS)* . Institute on Education and Training, RAND.

Koretz, D., Mitchell, K., Barron, S., and Keith, S. (1996). *Final* *report: Perceived effects of the Maryland school performance assessment program* [CSE Technical Report 409]. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST) (57 pages).

Kuhs, T., Porter, A., Floden, R., Freeman, D., Schmidt, W., and Schwille, J. (1985). Differences among teachers in their use of curriculum-embedded tests. *The Elementary School Journal*, 86(2), 141-153.

Lane, S. (1997, March). Framework for evaluating the consequences of an assessment program. Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

Lane, S., and Parke, C. (1996, April). Consequences of a mathematics performance assessment and the relationship between the consequences and student learning. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Langdon, C.A. (1997). The fourth Phi Delta Kappan poll of teachers' attitudes toward the public schools. *Phi Delta Kappan*, 79(3), 212-220.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis*, 15(1), 1-16.

Linn, R.L. (1994). Performance Assessment: Policy promises and technical measurement standards. *Educational Researcher*, 23(9), 4-14.

Linn, R.L., Baker, E.L., and Dunbar, S.B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15-21.

Linn, R.L., and Herman, J.L. (1997, February). *Standards-led assessment: Technical and policy issues in measuring school and student progress.* CSE Technical Report 426. National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Center for the Study of Evaluation, Graduate School of Education and Information Studies, University of California, Los Angeles.

Madaus, G. F. (1985). Test scores as administrative mechanisms in educational policy. *Phi Delta Kappan*, 66(9), 611-617.

Madaus, G.F. (1991). The effects of important tests on students: Implications for a National Examination System. *Phi Delta Kappan*, 73(3), 226-231.

Madaus, G.F., West, M.M., Harmon, M.C., Lomax, R.G. and Viator, K.A. (1992, October). The influence of testing on teaching math and science in grades 4-12. Executive Summary. National Science Foundation Study, Center for the Study of Testing, Evaluation, and Educational Policy, Boston College, Chestnut Hill, Massachusetts.

Mayes, M. (1997, August 30). Test results make school chief smile. "The Lansing State Journal," pp. 1A, 5A.

McDonnell, L. M. (1997). The politics of state testing: Implementing new student assessments [CSE Technical Report 424]. University of California, Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

McDonnell, L.M. and Choisser, C. (1997, September). Testing and teaching: Local implementation of new state assessments. CSE Technical Report 442. National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Center for the Study of Evaluation (CSE) Graduate School of Education and Information Studies, University of California, Los Angeles, CA.

Mehrens, W.A. and Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless or fraudulent? *Educational Measurement: Issues and Practices*, 8(1), 14-22.

Miller, M.D. (1998, February). Teacher uses and perceptions of the impact of statewide performance-based assessments. Council on Chief State School Officers. State Education Assessment Center, Washington, D.C.

Petrie, H.G. (1987). Introduction to "evaluation and testing." *Educational Policy*, 1, 175-180.

Pipho, C. (1997). Standards, assessment, accountability: The tangled triumvirate. *Phi Delta Kappan*, 78(9), 673-674.

Pomplun, M. (1997). State assessment and instructional change: A path model analysis. *Applied Measurement in Education*, 10(3), 217-234.

Porter, A. C., Floden, R. E., Freeman, D. J., Schmidt, W. H., and Schwille, J. P. (1986). Content determinants [Research Series No. 179]. Michigan State University, East Lansing, MI: Institute for Research on Teaching.

Rafferty, E. A. (1993, April). Urban teachers rate Maryland's new performance assessments. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Reckase, M. D. (1997, March). Consequential validity from the test developers' perspective. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Roeber, E., Bond, L.A., and Braskamp, D. (1997). Trends in statewide student assessment programs, 1997. North Central Regional Educational Laboratory and the Council of Chief State School Officers.

Rose, L.C., Gallup, A.M. and Elam, S.M. (1997). The 29th annual Phi Delta Kappa/Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan*, 79(1), 41-58.

Ryan, J.M. (1997). Personal communication.

Seidman, R. H. (1996, July 24). National education 'Goals 2000': Some disastrous unintended consequences. *Education Policy Analysis Archives*, 4, No. 11 (Entire Issue). (Available online at http://epaa.asu.edu/epaa/v4n11/).

SERVE. (1994). A new framework for state accountability systems [Special report of The Southeastern Regional Vision for Education].

Shepard, L. A., (1991). Will national tests improve student learning? *Phi Delta Kappan*, 72, 232-238.

Shepard, L. A., Flexer, R. J., Hiebert, E. H., Marion, S. F., Mayfield, V., and Weston, T. J. (1996, Fall). Effects of introducing classroom performance assessments on student learning. *Educational Measurement: Issues and Practices*, pp. 7-18.

Smith, M. L., Noble, A., Heinecke, W., Seck, M., Parish, C., Cabay, M., Junker, S., Haag, S., Tayler, K., Safran, Y., Penley, Y., and Bradshaw, A. (1997). Reforming schools by reforming assessment: Consequences of the Arizona student assessment program (ASAP): Equity and teacher capacity building [CSE Technical Report 425]. University of California, Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Smith, M.L., and Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice*, 10(4), 7-11.

Stecher, B.M. and Mitchell, K.J. (1995, April). Portfolio-driven reform: Vermont teachers' understanding of mathematical problem solving and related changes in classroom practice. CSE Technical report 400. National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Graduate School of Education and Information Studies, University of California, Los Angeles.

Stecklow, S. (1997, September 2). Kentucky's teachers get bonuses, but some are caught cheating. "The Wall Street Journal," pp. A1 and A5.

Stedman, L.C. (1996, January 23). The achievement crisis is real: A review of *The manufactured crisis. Education Policy Analysis Archives*, 4, No. 1 (Entire issue). (Available online at http://epaa.asu.edu/epaa/v4n1.html).

Taleporos, E. (1997, March). Consequential validity. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Womer, F.B. (1984). Where's the action? *Educational Measurement: Issues and Practice*, 3(3), 3.

Notes

1. This paper is a slight revision of the 1998 Vice Presidential address for Division D, American Educational Research Association, presented in San Diego. I would like to thank Bob Floden and Joe Ryan for helpful comments on a previous draft of this paper. Opinions expressed are those of the author and not necessarily those of the two reviewers.
2. As Jones has wondered "Can he really mean that?" (Jones, 1997, p. 3).
3. Space does not permit me to do justice to this very thorough report. I urge readers to obtain the report and study it carefully. Most of these state assessments equate through anchor items, and these items may not be totally secure.

About the Author

William A. Mehrens

Professor of Measurement
462 Erickson Hall
Michigan State University
East Lansing, MI 48824

517-355-9567
FAX 517-353-6393

WMehrens@pilot.msu.edu

WILLIAM A. MEHRENS is a professor of measurement at Michigan State University. He received his Ph.D. in educational psychology from the University of Minnesota in 1965. His interests include educational testing in general, legal issues in high-stakes testing, teaching to the test, and performance assessment. He has been elected to office in several professional organizations including the presidency of both the National Council on Measurement in Education (NCME) and the Association for Measurement and Evaluation in Guidance (currently called the Association for Assessment in Counseling). He is the immediate past Vice President of Division D of the American Educational Research Association (AERA). He is the author or co-author of several major textbooks and many articles. Honors include the NCME Award for Career Contributions to Educational Measurement, 1997; a University of Nebraska-Lincoln Teachers College Alumni Association Award of Excellence, 1997; an AACD Professional Development Award, 1991; MSU Distinguished Faculty Award, 1983; APA Division 15 Fellow, 1984; APA Division 5 Fellow, 1978; and Pi Mu Epsilon, National honorary mathematics fraternity, 1958.

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

Greg Camilli
Rutgers University

John Covaleskie
Northern Michigan University

Andrew Coulson
a_coulson@msn.com

Alan Davis
University of Colorado, Denver

Sherman Dorn
University of South Florida

Mark E. Fetler
California Commission on Teacher Credentialing

Richard Garlikov
hmwkhelp@scott.net

Thomas F. Green
Syracuse University

Alison I. Griffith
York University

Arlen Gullickson
Western Michigan University

Ernest R. House
University of Colorado

Aimee Howley
Marshall University

Craig B. Howley
Appalachia Educational Laboratory

William Hunter
University of Calgary

Richard M. Jaeger
University of North Carolina--Greensboro

Daniel Kallós
Umeå University

Benjamin Levin
University of Manitoba

Thomas Mauhs-Pugh
Rocky Mountain College

Dewayne Matthews
Western Interstate Commission for Higher Education

William McInerney
Purdue University

Mary P. McKeown
Arizona Board of Regents

Les McLean
University of Toronto

Susan Bobbitt Nolen
University of Washington

Anne L. Pemberton
apembert@pen.k12.va.us

Hugh G. Petrie
SUNY Buffalo

Richard C. Richardson
Arizona State University

Anthony G. Rud Jr.
Purdue University

Dennis Sayers
Ann Leavenworth Center
for Accelerated Learning

Jay D. Scribner
University of Texas at Austin

Michael Scriven
scriven@aol.com

Robert E. Stake
University of Illinois--UC

Robert Stonehill
U.S. Department of Education

Robert T. Stout
Arizona State University

archives | abstracts | editors | board | submit | comment | subscribe | search

EPAA
• enter the archives • browse the abstracts • the editors • the edit board
• submit article • submit commentary • search • subscribe
volume: • 1 • 2 • 3 • 4 • 5 • 6 • 7

This article has been retrieved 1952 times since July 21, 1998.

Education Policy Analysis Archives

Volume 6 Number 14 July 21, 1998 ISSN 1068-2341

A peer-reviewed scholarly electronic journal.
Editor: Gene V Glass Glass@ASU.EDU.
College of Education
Arizona State University,Tempe AZ
85287-2411

Copyright 1998, the EDUCATION POLICY ANALYSIS ARCHIVES.Permission is hereby granted to copy any article provided that EDUCATION POLICY ANALYSIS ARCHIVES is credited and copies are not sold.

Some Comments on Assessment in U.S. Education

Robert Stake
University of Illinois

Abstract We do not know much about what assessment has accomplished but we know it has not brought about the reform of American Education. The costs and benefits of large scale mandated achievement testing are too complex to be persuasively reported. Therefore, educational policy needs to be based more on deliberated interpretations of assessment, experience, and ideology. Evaluation of assessment consequences, however inconclusive, has an important role to play in the deliberations.

During the last half of the Twentieth Century in America, the traditional quality control of schooling, i.e., informal management (by teachers as well as administrators) board oversight, parent complaint, state guideline and regional accreditation, have continued to be prominent in school operations. But because the perceived quality of public education has fallen off, other means have been added to evaluate and to improve teaching and learning. For thirty years, assessment has been a significant means of quality control and instrument of educational reform.

Earlier, in the Century's third quarter, the impetus for changing American schooling was the appearance of Sputnik. It was reasoned that American schools were unsuccessful if the Soviets could be first to launch spacecraft. College professors and the National Science Foundation stepped forward to redefine mathematics education and the rest of the curriculum, creating a "new math," inquiry teaching, and many courses strange to the taste of most teachers and parents. According to Gallup polls year after year, citizens expressed confidence in the local school but increasingly worried about the national system. In the 1960s, curriculum redevelopment was the main instrument of reform but, in the 1970s, state-level politicians, reading the public as unhappy both with tradition and federalized reform, created a reform of their own. Their reform spotlighted assessment of student performance.

The term "assessment" then became taken to mean the testing of student achievement with standardized instruments. Student performance goals were made more explicit so that testing could be more precisely focused, and efforts were made to align curricula with the testing. Schooling includes many performances, provisions, and relationships which could be assessed but attention came down predominantly on the students: "If they haven't learned, they haven't been taught."

Now for at least two decades, in almost every school, at every grade level and in each of the subject matters, student achievement has been assessed. And every year, it has been found largely unchanged from previous testing. Over the same periods, teaching, on the whole, appears to have been little changed, certainly not restructured. Explication of goals appears not to have set more achievable targets. The last decade has seen efforts to set standards particularly for levels of student performance needed to restore American Education to a leading, world position. From time to time, gains occurred, but small and not sustained--losses also occurred. Instead of reading this lack of sustained progress as pointing to need for a different grand strategy, the clearest summons has been for additional assessment.

Purposes and Expectations of Assessment

Goal statements are simplifications. The felt purposes of education, aggregated across the profession, across researchers, the public and the primary beneficiaries, are far more complex than those represented in goal statements and formal assessments. Facts, theories, and reasoning are needed not just in isolation but interactively, innovatively, in a range of contexts. We hold a vast inventory of expectations, beyond catalogue, partly ineffable, often only apparent in disappointments as students fall short. That immense inventory is approximated by the informal assessments by teachers much better than by explicated lists of goals.

The grand manifold of purposes of Education held by any one person at any one time also is complex, and situational and internally contradictory. People, even those specially trained, are not very good at speaking of "what all they expect" of an educated person. Again, the complexity shows most forcefully when the person does not perform well. Any one shortfall tells little about the array of purposes. Any one assessment, however precise and valid, does not sample well the manifold of purposes. Broad and attentive use of assessments, formal and informal, evokes realization that what we expect of students and the uses to be made of a graduate's education extend far beyond formal goals, standards and lesson plans. Formal representations of aim and accomplishment provide flimsy accounts of the real thing.

This is not to suggest it useless to record educational purposes and student performance. It is useful to categorize them, to illustrate and prioritize them, sometimes by abilities and subject matters--but always a risk. The subsets or domains are artificial. Needed in the anticipation and provision of Education, they often serve poorly to represent the education a student is attaining. Assessment based strongly on goals or domains is likely to tell more about the territory of teaching than the territory of learning.

Procedurally, Education is organized at the level of courses and classrooms, then lessons and assessments. Actually, education occurs in complex and differentiated ways in each child's mind. Assessments tuned to management levels cannot be expected to mirror the complexity of learning and diversity of learners. However carefully named and designed, mean scores do not necessarily indicate basic accomplishments for a group of learners. Each testing needs empirical validation.

Validation of Assessment

Standardized test development is one of the most technically sophisticated specialties within Education. Definitions and analytic procedures, at least at the major testing companies are scrutinized, verified, codified and reworked. The traditional ethics of psychometrics call for extensive construct validation of the measurements to be used in schooling. And it is not enough that the instruments and operations be examined for accuracy, relevance and freedom from bias, but that independent measurements be used to confirm that scores indicate what we think they indicate. Sound test development is a slow and expensive procedure.

In the development of assessment instruments by the 50 states, adequate validation has seldom taken place. Instruments have been analyzed statistically to see that they are internally consistent but not that mean what users think they mean. Presumption that assessments indicate quality of teaching, appropriateness of curricula, and progress of the reform movement-- commonplace presumptions in political and media dialogue--is unwarranted. Proper validation would tell us the strength or weakness of our conclusions about student accomplishment. Those studies have not been commissioned. The most needed validation of statewide assessment programs has not taken place.

The question of whether or not the assessment legislation, as opposed to the assessment scores, is having a good effect on student education is a separate question. Assessment changes instruction. Reformists expect assessment will force teachers to teach differently, and, in various ways and to various extents, they do. Each assessment effort will have both positive and negative consequences. The design and promulgation of an assessment program is only an approximation of what actually occurs. The operation described in any report is a partial misrepresentation of institutional initiative and measurement integrity. For a reader, it is an opportunity to misperceive what is happening in the schools and the lives of youngsters. We need better descriptions, better evidence, of those consequences of assessment. And partly because we construct nuances of meaning faster than we invent measurements, we need to understand that we will never have a clear enough picture of the consequences of assessment. All findings should be treated as partial and tentative.

Value Determination

Not only has there been an increase in the amount of formal educational assessment but assessment has been applied increasingly to influence the well-being of students, schools and systems. The "stakes" have risen. Funding, autonomy and privilege have been attached to levels of scoring. The intention has been to get students and teachers dedicated to their tasks, and this sometimes happens, but there have been costs as well as benefits. Among the reported negative consequences of raising the stakes of assessment are:

- instruction is diverted,
- student self-esteem is eroded,
- teachers are intimidated,
- the locus of control of education is more centralized,
- undue stigma is affixed to the school,
- school people are lured towards falsification of scores,
- some blame for poor instruction is redirected toward students when it should rest with the profession and the authorities, and
- the withholding of needed funding for education appears warranted.

The most obvious consequence of increased assessment is that teachers increase preparation for test taking, including test- taking skills and greater familiarization with the anticipated content of testing. Also, topics tested are considered of higher priority and topics untested slip in priority. Assessments are not diagnostic. There is little strategic theory fitting pedagogy to assessment so that few teachers know how to respond to poor student performance, other than to try harder. Thus, over-emphasis on assessment erodes confidence in legitimate teaching competence.

As the stakes rise, the central authorities are both pressured and authorized to intervene more in teaching responsibilities. A widespread public perception of legislators and school authorities is that they are not knowledgeable or competent in matters of the classroom. With ever-confirming evidence that students continue to be testing poorly, the public is tempted to withhold funds for needed improvement in instruction. There is good evidence that increased funding alone will not greatly change the quality of teaching. But at the same time, by investing in the assessment of students without investing in more direct evaluation of teacher and administrative performance, the professional people and the elected overseers are partly "off the hook." In summary, the consequences of assessment are complex, extending far beyond the redirecting of instruction toward state goals.

It is too much to expect that we soon will clearly discern the consequences of assessment and, even less soon, what caused them. Both the consequences and the causes are complex, both as to constituents and as to conditions. Lacking an adequate research base, curricular policy needs to be based on deliberations, long and studied interpretation of assessment, experience, and ideology. That is unlikely when professional wisdom is getting little respect. Often the public presumes that educators put their own interests above those of students. But good deliberations are not uncommon. Evaluation of the consequences of assessment has an important role informing those deliberations.

Even if we were able to improve determination of the consequences of assessment, we lack theory and management systems that guide us in applying that information to the improvement of teaching and learning. We need not wait for politics or the professional to be reformed. We can rely on the political, intuitive, and leadership processes we now have to make assessment more a positive and less a negative force within education.

As indicated before, people do have different purposes for education and for assessment. And for any one purpose, they value the results differently. That is just part of the reality, neither excusing nor facilitating the assessment of assessment.

The assessment practice that does the most measurable, immediate good is not necessarily the practice that has the best long range effect. For example, using testing time entirely for easily measured skills instead of partly for "ill-defined" interpretive experience increases precision and predictive validity but discourages well-thought-out advocacies to include problem-solving experience throughout elementary school. Value trade-offs need to be considered for long-term as well as short- term effects.

Curriculum and Instruction

Management of teaching and the curriculum cannot be effective without assessment. The best and the worst assessment we have is informal and teacher-driven, sometimes capricious and sometimes more aimed at avoiding embarrassment than maximizing services to children. Yet, it works pretty well, sensitive to what individual children are doing, viewed favorably by a substantial proportion of parents and citizens, especially those people who interact themselves, even in small ways, with the academic program. Still, instructional assessment could be much, much better, and too little professional development is so aimed. The present informal assessment system is little engaged with the formal management information system of school districts and even less with the state's student achievement testing apparatus.

The most successful school improvement efforts have been those that decentralize and protect authority so that a match can be made between what the teachers want to teach and the parents and immediate community want taught. The present decade's "standards movement" was a step in the wrong direction, a further imposition of external values. Assessment was used to nullify decentralization efforts. The state does have a stake in what every child is learning but the state is poorly served by having each child trying to learn the same things. Accountability of the schools is in no way dependent on having each child tied to a core curriculum and tested on the same items. A single test for all is cheaper, but not a service to a diverse population of children.

State assessment is not wrong in its most general finding that teaching and learning in the American schools are mediocre. And that the range across districts is huge. The spread of achievement scores is stable and predictable, more a function of a child's lifetime educational opportunity than of what happens during a year in a classroom. Neither massive changes at home or in the classroom are likely to result in substantial gains on current assessment instruments.

As stated earlier, the validity of measurement of achievement is not the same as validity of those same scores as an indicator of quality of teaching and learning conditions. Teaching can be changed in a number of important ways within a school or classroom without change in achievement means. Using those scores as a measure of school improvement has not been validated. No accumulation of evidence shows assessment to be an indicator of good schooling. In spite of the absence of validity, assessment means continue to be the primary criterion for reform in a vast number of school districts. Given vigorous school improvement efforts over 20-30 years within countless districts, essentially all of them unaccompanied by substantial change in assessment results, what should be concluded is that testing is insensitive to important changes in teaching or that schools cannot be improved. The latter is untenable.

Uses and Stakes

The uses to which assessment information will be put varies not just across assessment approaches but greatly within approaches as well. Different school systems, teachers, and children, even those greatly alike, will be affected differently. It is not reasonable to suppose that the stakes of assessment are unimportant if they have little impact upon the majority. Special attention needs to be given to how assessment consequences affect the least privileged families and most vulnerable children.

One of the primary stakes of testing is the well-being of teachers. Teachers have much to lose in a high stakes assessment system. Assessment should not be avoided just because teachers protest but their working conditions and professional wisdom should not be trivialized. Teaching quality should be scrutinized. Student performance should be considered but it should not be a primary determinant of teaching competence. There is only a small connection between how well a teacher teaches and how well a child performs on a test.

One of the consequences of high stakes testing is the manipulation of rosters to excuse poor scoring children from participation. The most common way at present appears to be to have children classified as "special education" students, but a good bit of ingenuity has been shown in optimizing rosters.

High stakes assessment often does result in raised scores but the validity of widespread gains, locally or across the country, has not been established. No one wants to challenge the gains that appear, but presently emphasis on small changes serves to orient the school to the assessments rather than to education. Many of the consequences of assessment are best learned from the people who administer the tests, even though they have a self- interest. Many are quick to acknowledge that the assessment enterprise is flawed.

Good research can help but it is mostly a professional and political matter. Until community attitude sets out to make the best of the schools, less to blame them, (however much they deserve the blame), not much good will happen. This is not a nation dedicated to the best possible education system. There are lots of people who would rather have lower taxes than to extend educational benefits. Higher taxes do not assure better opportunities but an interest in finding better opportunities is not a national purpose. Looking at it simplistically, support for assessments appears to be a step toward improving education, but the quarter-century record shows that assessment-driven reform has not worked. Why does it continue to be politically popular? The main consequence of assessment-based reform is that education has not substantially improved. We do not lack evidence of that.

About the Author

Robert E. Stake
University of Illinois--Urbana, Champaign

Email: r-stake@uiuc.edu

Robert Stake is professor of education and director of CIRCE at the University of Illinois. Since 1963 he has been a specialist in the evaluation of educational programs, moving from psychometric to qualitative inquiries. Among the evaluative studies he has directed are works in science and mathematics in elementary and secondary schools, model programs and conventional teaching of the arts in schools, development of teaching with sensitivity to gender equity; education of teachers for the deaf and for youth in transition from school to work settings, environmental education and special education programs for gifted students, and the reform of urban education. Stake has authored *Quieting Reform*, a book on Charles Murray's evaluation of Cities-in -Schools; two books on methodology, *Evaluating the Arts in Education* and *The Art of Case Study Research*; and *Custom and Cherishing*, a book with Liora Bresler and Linda Mabry on teaching the arts in ordinary elementary school classrooms in America. Recently he led a multi-year evaluation study of the Chicago Teachers Academy for Mathematics and Science. For his evaluation work, in 1988, he received the Lazarsfeld Award from the American Evaluation Association, and, in 1994, an honorary doctorate from the University of Uppsala.

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

John Covaleskie
Northern Michigan University

Alan Davis
University of Colorado, Denver

Mark E. Fetler
California Commission on Teacher Credentialing

Thomas F. Green
Syracuse University

Arlen Gullickson
Western Michigan University

Aimee Howley
Marshall University

William Hunter
University of Calgary

Daniel Kallós
Umeå University

Thomas Mauhs-Pugh
Rocky Mountain College

William McInerney
Purdue University

Les McLean
University of Toronto

Anne L. Pemberton
apembert@pen.k12.va.us

Richard C. Richardson
Arizona State University

Dennis Sayers
Ann Leavenworth Center for Accelerated Learning

Michael Scriven
scriven@aol.com

Robert Stonehill
U.S. Department of Education

Greg Camilli
Rutgers University

Andrew Coulson
a_coulson@msn.com

Sherman Dorn
University of South Florida

Richard Garlikov
hmwkhelp@scott.net

Alison I. Griffith
York University

Ernest R. House
University of Colorado

Craig B. Howley
Appalachia Educational Laboratory

Richard M. Jaeger
University of North Carolina--Greensboro

Benjamin Levin
University of Manitoba

Dewayne Matthews
Western Interstate Commission for Higher Education

Mary P. McKeown
Arizona Board of Regents

Susan Bobbitt Nolen
University of Washington

Hugh G. Petrie
SUNY Buffalo

Anthony G. Rud Jr.
Purdue University

Jay D. Scribner
University of Texas at Austin

Robert E. Stake
University of Illinois--UC

Robert T. Stout
Arizona State University

---

archives | abstracts | editors | board | submit | comment | subscribe | search

• enter the archives • browse the abstracts • the editors • the edit board
• submit article • submit commentary • search • subscribe
volume: • 1 • 2 • 3 • 4 • 5 • 6 • 7

This article has been retrieved 893 times since August 19, 1998.

Education Policy Analysis Archives

Volume 6 Number 15 **August 19, 1998** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal.
Editor: Gene V Glass Glass@ASU.EDU.
College of Education
Arizona State University,Tempe AZ
85287-2411

Copyright 1998, the EDUCATION POLICY ANALYSIS ARCHIVES.Permission is hereby granted to copy any article provided that EDUCATION POLICY ANALYSIS ARCHIVES is credited and copies are not sold.

A Note on the Empirical Futility of Labor-Intensive Scoring Permutations for Assessing Scholarly Productivity: Implications for Research, Promotion/Tenure, and Mentoring

Christine Hanish
John J. Horan
Bethanne Keen
Ginger Clark

Arizona State University

Abstract The measurement of scholarly productivity is embroiled in a controversy concerning the differential crediting of coauthors. Some researchers assign equivalent shares to each coauthor; others employ weighting systems based on authorship order. Horan and his colleagues use simple publication totals, arguing that the psychometric properties of labor-intensive alternatives are unknown, and relevant ethical guidelines for including coauthors are neither widely understood nor consistently followed. The PsycLIT and SSCI data bases provided exhaustive publication and citation frequencies for 323 counseling psychology faculty. All PsycLIT scoring permutations yielded essentially identical information; inter-correlations ranged from .96 to unity. Moreover, all PsycLIT methods correlated highly with SSCI within a very narrow band. Since attention to the number and/or ordinal position of coauthors yields no useful information, productivity should be defined parsimoniously in terms of simple publication counts. Implications for research, promotion/tenure, and the mentoring of graduate students are discussed.

Publishing behavior is perhaps the most revered and reviled variable in education and psychology. The bipolar affect it generates undoubtedly derives from the fact that although the act of publishing is inextricably entwined with status and the reward system of a scientific discipline (e.g., promotion, tenure, merit pay, and the like), the criteria for evaluating what an individual publishes are much less clear (Merton, 1973). The concepts of *productivity*, *impact*, and *quality* are often used interchangeably as descriptors, yet there are important methodological and psychometric differences.

Productivity refers to the quantity of publications attributable to a given scholar, expressed as a lifetime total or a yearly rate when divided by the scholar's professional age. Impact generally means how frequently that individual's work is cited by other authors, which likewise can be expressed as a lifetime total or a yearly rate. Quality is almost never assessed directly; productivity and impact, though, frequently pose in its place (see Keen, Horan, Hanish, Copperstone, & Tribbensee, 1998).

Since vita entries provide no assurance that a document really exists, the assessment of productivity is usually confined to the number of publications by an individual that appear in large data-bases such as ERIC or PsycLIT (Horan & Erickson, 1991). The gate-keeper functions in these data-bases, however, infuse raw counts of productivity with elements of quality. For example, PsycLIT only lists articles that appear in refereed journals recognized by APA as relevant to the discipline of psychology.

The assessment of impact is likewise usually restricted to full citation histories contained in large holdings such as SSCI, though smaller segments of that data base and/or fewer numbers of outlets have been used (albeit, unreliably, see Horan, Hanish, & Beasley, 1995). SSCI is more often associated with quality than is PsycLIT, but that kudo may not be warranted. Hanish, Horan, Keen, St. Peter, Ceperich, and Beasley (1995) reported high relationships between PsycLIT and SSCI; moreover, other limitations of SSCI are less well known and understood. For example, SSCI scores may be inflated by hidden self-citations, citations by prolific colleagues, advisees, or significant others, the notoriety of a study rather than its importance, and so forth (see Horan, Hanish, Keen, Saberi, & Hird, 1993).

The measurement of productivity has become embroiled in a controversy concerning the differential crediting of coauthors. Some researchers (such as Bohn, 1966; Goodstein, 1963; Goodyear, Abadie, & Walsh, 1983; Katz & Brophy, 1975; Tinsley & Tinsley, 1979; Walsh, Feeney, & Resnick, 1969) give each coauthor equal partial credit (e.g., a third of a point to three coauthors of a given article); others (such as Delgado & Howard, 1994; Ellis, Haase, Skowron, & Kaminsky, 1993; Howard, 1983; Howard, Cole, & Maxwell, 1987; Osipow, 1985; Skovholt, Stone, & Hill, 1984) apply various weighting formulas based on the ordinal positions of coauthors (e.g., first author receives half of the credit, the second author 30% of the credit, and the last author the final 20%).

In contrast, Horan and his colleagues (e.g., Hanish, et al.,1995; Horan & Erickson, 1991; Horan, Weber, Fitzsimmons, Maglio, & Hanish, 1993b) have always used simple raw PsycLIT totals for each author, arguing that the psychometric properties of the foregoing schema are unknown, and APA's ethical guidelines for assigning authorship are neither widely understood nor consistently followed (e.g., see Fine & Kurdek, 1993; Goodyear, Crego, & Johnston, 1992).

The present study, therefore, attempted to clarify the relationships between the various scoring permutations of PsycLIT with each other and with SSCI. Although the same scoring controversy could apply to coauthorships listed in ERIC or in other data bases, we chose PsycLIT because its refereed holdings are obtained independent of author consent, and thus provide a more meaningful basis for comparison with other indices of scholarly merit.

Method

Subjects

Hanish et al. (1995), identified the entire population of academic counseling psychology faculty (*n* = 323) who were members of Division 17 and who had governance responsibilities in any active doctoral training program; for each individual, they secured complete PsycLIT data from 1974 to 1991 and SSCI data from 1971 to 1991. In the present study we updated all PsycLIT and SSCI data on these individuals to be current to 1996.

Measures

The PsycLIT data base includes all *Psychological Abstracts* references attributable to individual authors published from 1974 to present. A search by author name yielded a full bibliographical citation list for that author including coauthors and abstracts. These data were scored according to six different methods described as follows:

- *Method 1*, used by Horan and his associates (e.g., Horan & Erickson, 1991; Hanish et al., 1995), awards a single point to each author for each publication regardless of the number of coauthors or their ordinal position. If an individual has 13 sole or coauthored publications in the PsycLIT data base his or her score will be 13.

- *Method 2* is relatively popular (e.g., Bohn, 1966; Goodstein, 1963; Goodyear, Abadie, & Walsh, 1983; Katz & Brophy, 1975; Tinsley & Tinsley, 1979; Walsh, Feeney, & Resnick, 1969); coauthors receive equal partial credit (e.g., a third of a point to three coauthors of a given article). First and last authors are treated alike. Method 2 and all methods that follow are increasingly labor intensive in that they require the computation and summing of various amounts of credit for each bibliographic entry on a given author's publication record.

- *Method 3* (Delgado & Howard, 1994; Howard, 1983) awards one point to sole authors. The first and second authors of a coauthored publication receive .67 and .33 points, respectively. If three coauthors are involved, the differential credit allocations are .50, .30, and .20. Additional coauthors result in decreasing credit for all.

- *Method 4* (Howard, Cole, & Maxwell, 1987) uses a very complex formula to compute the differential allocation of credit. As with Method 3, authors and coauthors receive declining amounts of credit as their numbers increase and their ordinal positions descend.

- *Method 5* (Osipow, 1985; Skovolt, Stone, & Hill, 1984) awards sole authors and first authors 5 points, second authors 4, third authors 3, and fourth authors 2; all subsequent coauthors receive a score of 1. Points are thus constant across ordinal position.

- *Method 6* was devised by Ellis, Haase, Skowron, and Kaminsky (1993). Weights depend on the number of authors, the order of authorship, and the value of the article using the method of Skovolt, Stone, and Hill (1984). For example, an article with three coauthors has a value of 12 which is derived by adding five points for the first author, four points for the second author, and three points for the third author. The first author's credit then is 5/12 or .417; the second author's credit is 4/12 or .333 and so on. For articles with more than four coauthors, the fifth and subsequent authors receive equal shares of .067 such that, for example, the fifth and sixth authors would each receive .034.

The credit consequences of the six different productivity scoring methods on the coauthors of a given article can be seen in Table 1.

Table 1

Template for Productivity Scoring Methods Indicating Comparative Credit by Number and Ordinal Position of Coauthors.

Author/ Coauthors	Method 1 Horan	Method 2 Walsh	Method 3 Howard 1	Method 4 Howard 2	Method 5 Skovholt	Method 6 Ellis
1/1	1.000	1.000	1.000	1.00	5.000	1.000
1/2	1.000	.500	.670	.600	5.000	.556
2/2	1.000	.500	.330	.400	4.000	.444
1/3	1.000	.333	.500	.474	5.000	.417
2/3	1.000	.333	.300	.316	4.000	.333
3/3	1.000	.333	.200	.210	3.000	.250
1/4	1.000	.250	.400	.415	5.000	.357
2/4	1.000	.250	.300	.277	4.000	.286
3/4	1.000	.250	.200	.185	3.000	.214
4/4	1.000	.250	.100	.123	2.000	.143
1/5	1.000	.200	.330	.384	5.000	.333
2/5	1.000	.200	.270	.256	4.000	.267
3/5	1.000	.200	.200	.171	3.000	.200
4/5	1.000	.200	.130	.114	2.000	.133
5/5	1.000	.200	.070	.076	1.000	.067
1/6	1.000	.167	.286	.365	5.000	.333
2/6	1.000	.167	.238	.244	4.000	.267
3/6	1.000	.167	.190	.162	3.000	.200
4/6	1.000	.167	.143	.108	2.000	.133
5/6	1.000	.167	.095	.072	1.000	.035
6/6	1.000	.167	.048	.048	1.000	.035
1/7	1.000	.143	.250	.354	5.000	.333
2/7	1.000	.143	.214	.236	4.000	.267
3/7	1.000	.143	.179	.157	3.000	.200
4/7	1.000	.143	.143	.105	2.000	.133
5/7	1.000	.143	.107	.070	1.000	.023
6/7	1.000	.143	.071	.047	1.000	.023
7/7	1.000	.143	.036	.031	1.000	.023
1/8	1.000	.125	.222	.347	5.000	.333
2/8	1.000	.125	.194	.231	4.000	.267
3/8	1.000	.125	.167	.154	3.000	.200
4/8	1.000	.125	.139	.103	2.000	.133
5/8	1.000	.125	.111	.069	1.000	.017
6/8	1.000	.125	.083	.046	1.000	.017
7/8	1.000	.125	.056	.030	1.000	.017
8/8	1.000	.125	.028	.020	1.000	.017

Note: The names are those of researchers most closely associated with the various scoring methods. Under Author/Coauthors, 1/1 = sole author, 1/2 = first author of an article by two authors, 2/3 = second author of an article by three authors, etc.

SSCI is a compilation of citations to a given sole or first author by that same author and other scholars from 26 disciplines in the social and behavioral sciences. Cited authors are arranged alphabetically in bound volumes covering the years 1966 to present. Our search was confined to the SSCI volumes paralleling our PsycLIT database. Below each cited author's work in SSCI is a list of individuals who referenced that work along with abbreviated outlet information. We used two SSCI scoring methods, namely, the grand total and the grand total minus obvious self-citations. An obvious self-citation occurred when a first author cited himself or herself in a first-authored reference. SSCI makes no provision for detecting "hidden" self-citations, for example, second authors citing their first-authored works.

Procedures

Procedures for faculty identification, biographical information, reliability analyses, and so forth are described in Hanish et al. (1995). The new PsycLIT and SSCI raw data obtained for the present study were secured in the same fashion. Each of the 323 faculty publication histories was then coded according to the methods described above by doctoral students working independently. This, of course, was an extremely time-consuming process. A random sample of 1752 publications was rechecked by additional students; disagreements between coders were trivial (1.9%). To facilitate further work in this area, *a priori* scoring templates are presented in Table 1. For example, if an individual is listed as third of four authors on a particular publication, the columns contain the precalculated author-position scores for each of the six methods.

Results

The actual raw data on which all analyses are based are being made available to the reader. From this point, the data files can be accessed in EXCEL, SPSS or ACII format. Of 323 individual faculty, only 10 had no evidence of publishing history in the PsycLIT and SSCI data bases. A similar number exceeded 65 publications and 650 citations. The median faculty member in our study had 13 publications in PsycLIT and was cited in SSCI 50 times including an average of 3 obvious self-citations. Table 2 depicts the correlations involving PsycLIT scoring permutations with each other and with SSCI.

Table 2

Correlations between PsycLIT and SSCI scoring permutations

Variable	PsycLIT Method 2 Walsh	PsycLIT Method 3 Howard1	PsycLIT Method 4 Howard2	PsycLIT Method 5 Skovholt	PsycLIT Method 6 Ellis	SSCI Total	SSCI Minus SelfCites
PsycLIT Method 1 Horan	.961	.963	.965	.998	.966	.711	.669
PsycLIT Method 2 Walsh		.997	.998	.971	.999	.703	.659
PsycLIT Method 3 Howard1			1.00	.975	.999	.701	.654
PsycLIT Method 4 Howard2				.976	1.00	.703	.657
PsycLIT Method 5 Skovholt					.976	.712	.669
PsycLIT Method 6 Ellis						.704	.659
SSCI Total							.995

Note: The names are those of researchers most closely associated with the various scoring methods.

The relationships among the six scoring methods for assessing productivity are remarkably high. No individual pairwise correlation was lower than .96; several r's reached unity. Similarly, the Pearson r between SSCI total and SSCI minus obvious self-citations also approached unity (.995).

More importantly, however, despite the fact that productivity and impact reflect different concepts and derive from disparate assessment methodologies, the relationships between these variables, regardless of scoring method, were strong and consistent. All six PsycLIT scoring permutations correlated with SSCI total inside a very narrow band of .701 to .712; and the band remained high and narrow (.654 to .669) when obvious self-citations were deleted.

Discussion

Our data reflect the lifetime publishing behavior of an entire population of academic faculty affiliated with doctoral training programs in counseling psychology. Although we have not established that the foregoing relationships hold true in other sectors of science, there are no *a priori* reasons to think otherwise. Essentially, the controversy involving the comparative merits of various methods for assessing scholarly productivity has been settled. All PsycLIT scoring permutations yield essentially identical information; inter-correlations range from .96 to unity. Moreover, all of these PsycLIT methods also correlate with SSCI data at a fairly high level and within a very narrow band.

Several implications are apparent. For example, future researchers are now informed that labor-intensive scoring permutations are not cost beneficial in comparison to the use of simple raw scores to assess an individual's scholarly productivity. The law of parsimony demands that a scholar's productivity be defined in terms of the number of articles carrying his or her name; attention to the number and/or ordinal position of coauthors yields no useful information.

It would be interesting to observe if the behavior of promotion and tenure committees will change as a result of increased awareness of the relationships reported in this study. Such committees can exhibit highly variable judgment even within the same institution. Collaborative research, for example, is sometimes valued ("has good collegial relationships"), sometimes denigrated ("needs to demonstrate more independent scholarship"); our findings suggest that the phenomenon of coauthoring is simply a facet of academic life, not a basis for evaluation.

Finally, we hope that our data eliminate a thorny disincentive to the formation of good mentoring relationships. Scoring methods 2 through 6 clearly advantage those in differential power relationships who chose self-interest over propriety while still staying within the letter of relevant ethical codes. Reptilian supervision modes are predictable, though no less abhorrent in the context of promotion, tenure, and merit pay systems that, for example, heavily weight sole authorships. Half of the publications by our institution's counseling psychology faculty in the PsycLIT data base involve students as coauthors, a percentage possibly comparable to that displayed in many other graduate programs. In contrast to labor-intensive, and empirically unwarranted alternatives, the use of simple raw scores to assess productivity contributes to the class-action benefit of everyone at no cost to anyone.

References

Bohn, M. J. (1966). Institutional sources of articles in this journal of counseling psychology--Four years later. *Journal of Counseling Psychology, 13,* 489-490.

Delgado, E. A., & Howard, G. S. (1994). Changes in research productivity in counseling psychology: Revisiting Howard (1983) a decade later. *Journal of Counseling Psychology, 41,* 69-73.

Ellis, M. V., Haase, R.F., Skowron, E. A., & Kaminsky, L. (1993, August). *Institutional affiliations of contributors to scholarly and professional activities in counseling psychology: 1987-1990.* Paper presented at the Annual Meeting of the American Psychological Association, Toronto, Canada.

Fine, M. A., & Kurdek, L. A. (1993). Reflections on determining authorship credit and authorship order on faculty-student collaborations. *American Psychologist, 48,* 1141-1147.

Goodstein, L. D. (1963). The institutional sources of articles in the Journal of Counseling Psychology. *Journal of Counseling Psychology, 10,* 94-95.

Goodyear, R. K., Abadie, P. D., & Walsh, W. B. (1983). Graduate school origins of Journal of Counseling Psychology authors: Volumes 15-28. *Journal of Counseling Psychology, 30,* 283-286.

Goodyear, R. K., Crego, C. A., Johnston, M. W. (1992). Ethical issues in the supervision of student research: A study of critical incidents. *Professional Psychology: Research and Practice, 23,* 203-210.

Hanish, C., Horan, J. J., Keen, B., St. Peter, C. C., Ceperich, S. D., & Beasley, J. F. (1995). The scientific stature of counseling psychology training programs: A still picture of a shifting scene. *The Counseling Psychologist, 23,* 82-101.

Horan, J. J., & Erickson, C. D. (1991). Fellowship behavior in Division 17 and the MOMM cartel. *The Counseling Psychologist, 19,* 253-259.

Horan, J. J., Hanish, C., & Beasley, J. F. (1995). A methodological reply to a motivational charge. *The Counseling Psychologist, 23,* 125-128.

Horan, J. J., Hanish, C., Keen, B., Saberi, D., & Hird, J. S. (1993a). When examining the cerebral functioning of Division 17, which organ should we dissect? *The Counseling Psychologist, 21,* 307-315.

Horan, J. J., Weber, W. L., Fitzsimmons, P., Maglio, C. J., & Hanish, C. (1993b). Further manifestations of the MOMM phenomenon: Relevant data on editorial board appointments and membership composition. *The Counseling Psychologist, 21,* 278-287.

Howard, G. S. (1983). Research productivity in counseling psychology: An update and generalization study. *Journal of Counseling Psychology, 30,* 600-602.

Howard, G. S., Cole, D.A., & Maxwell, S. E. (1987). Research productivity in psychology based on publication in the journals of the American Psychological Association. *American Psychologist, 42,* 975-986.

Katz, G. M. & Brophy, A. L. (1975). Institutional sources of articles in the Journal of Counseling Psychology, 1962-1973. *Journal of Counseling Psychology, 22,* 160-163.

Keen, B., Horan, J. J., Hanish, C., Copperstone, J., Tribbensee, N. (August, 1998). *Publication frequency, citation frequency, and quality of counseling psychology research*. Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. (N. W. Storer, Ed.). Chicago: The University of Chicago Press.

Osipow, S. H. (1985). Skovholt, Stone, and Hill's (1984) "Institutional affiliations of contributors to scholarly and professional activities in counseling psychology: 1980-1983" -- A critique. *Journal of Counseling Psychology, 32,* 466-468.

Skovholt, T.M., Stone, G. L., & Hill, C.E. (1984). Institutional affiliations of contributors to scholarly and professional activities in counseling psychology: 1980-1983. *Journal of Counseling Psychology, 31,* 394-397.

Tinsley, D. J. & Tinsley, H. E. A. (1979). Trends in institutional contributions to the Journal of Counseling Psychology. *Journal of Counseling Psychology, 26,* 152-158.

Walsh, W.B., Feeney, D., & Resnick, H. (1969). Graduate school origins of Journal of Counseling Psychology authors. *Journal of Counseling Psychology, 16,* 375-376.

About the Authors

Christine Hanish

Homepage: http://psy.ed.asu.edu/~hanish/

Christine Hanish is a doctoral student in counseling psychology at Arizona State University. She works for ASU's Preventive Intervention Research Center which specializes in the development and validation of programs for children, adolescents, and family. She is currently immersed in a research project attempting to establish the norms of scholarly behavior for academic counseling psychologists.

John J. Horan

Email: horan@asu.edu
Homepage: http://psy.ed.asu.edu
Personal Homepage: http://seamonkey.ed.asu.edu/~horan

I am a professor of counseling psychology at Arizona State University. I graduated from Michigan State University and taught at Penn State before moving to ASU in 1985. Most of my writing has focused on the evaluation of cognitive-behavioral intervention strategies.

For more than a decade I have been examining the experimental construct validity of these interventions. For example, do they produce changes on measures of high theoretical relevance while simultaneously failing to effect changes on measures of low theoretical relevance? Lately, I have concentrated on adapting and evaluating computer and Internet interventions for a variety of counseling problems.

For a quick look at how I squandered my youth, click on my web-based vita. My most important accomplishments, however, are not listed there. I have had many extraordinary students in my career, including those who share this masthead. I feel privileged to have contributed to their professional development; they surely have enhanced my own.

Bethanne Keen, Ph.D.
Email: BethKeen@aol.com
Homepage: http://psy.ed.asu.edu/~keen/

Bethanne Keen received a Ph.D. in counseling psychology from Arizona State University in December 1997. She is currently completing a postdoctoral residency in psychology with a large group practice in Phoenix, Arizona. She also serves as chair of the Legislative Affairs Committee for the Arizona Psychological Association. Her dissertation, currently being prepared for publication, explores the relationships between publication frequency, citation frequency and quality of research conducted by counseling psychologists in academe. She is currently involved in a research project designed to illuminate the challenges faced by new Ph.D.s in psychology to achieve employment and licensure in Arizona. Her other research interests include collection and analysis of clinical outcomes data.

Ginger Clark
ginger.clark@asu.edu
Homepage: http://psy.ed.asu.edu/~clark/

Ginger Clark is a doctoral student in counseling psychology at Arizona State University. She has conducted or contributed to studies in human sexual styles, parent education, parent education in career development, health habits, and quality of life for mid-life women. She has also written book reviews in the area of family therapy. Clark received her Bachelor's and Master's degrees in psychology at California State University Long Beach. She is currently in her fourth year of doctoral study, and is working toward an academic position in counseling psychology.

Correspondence concerning this article should be addressed to John J. Horan, Division of Psychology in Education, Arizona State University, BOX 870611, Tempe, AZ 85287-0611. Electronic mail may be sent to: horan@asu.edu

---

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

Greg Camilli
Rutgers University

John Covaleskie
Northern Michigan University

Andrew Coulson
a_coulson@msn.com

Alan Davis
University of Colorado, Denver

Sherman Dorn
University of South Florida

Mark E. Fetler
California Commission on Teacher Credentialing

Richard Garlikov
hmwkhelp@scott.net

Thomas F. Green
Syracuse University

Alison I. Griffith
York University

Arlen Gullickson
Western Michigan University

Ernest R. House
University of Colorado

Aimee Howley
Marshall University

Craig B. Howley
Appalachia Educational Laboratory

William Hunter
University of Calgary

Richard M. Jaeger
University of North Carolina--Greensboro

Daniel Kallós
Umeå University

Benjamin Levin
University of Manitoba

Thomas Mauhs-Pugh
Rocky Mountain College

Dewayne Matthews
Western Interstate Commission for Higher Education

William McInerney
Purdue University

Mary P. McKeown
Arizona Board of Regents

Les McLean
University of Toronto

Susan Bobbitt Nolen
University of Washington

Anne L. Pemberton
apembert@pen.k12.va.us

Hugh G. Petrie
SUNY Buffalo

Richard C. Richardson
Arizona State University

Anthony G. Rud Jr.
Purdue University

Dennis Sayers
Ann Leavenworth Center
for Accelerated Learning

Jay D. Scribner
University of Texas at Austin

Michael Scriven
scriven@aol.com

Robert E. Stake
University of Illinois--UC

Robert Stonehill
U.S. Department of Education

Robert T. Stout
Arizona State University

---

archives | abstracts | editors | board | submit | comment | subscribe | search

EPAA
• enter the archives • browse the abstracts • the editors • the edit board
• submit article • submit commentary • search • subscribe
volume: • 1 • 2 • 3 • 4 • 5 • 6 • 7

This article has been retrieved 1997 times since August 20, 1998.

Education Policy Analysis Archives

Volume 6 Number 16 **August 20, 1998** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal.
Editor: Gene V Glass Glass@ASU.EDU.
College of Education
Arizona State University,Tempe AZ
85287-2411

Copyright 1998, the EDUCATION POLICY ANALYSIS ARCHIVES.Permission is hereby granted to copy any article provided that EDUCATION POLICY ANALYSIS ARCHIVES is credited and copies are not sold.

Criticizing the Schools: Then and Now

Benjamin Levin
The University of Manitoba

Abstract Schools in many countries are facing intense and elevated levels of criticism, with much debate over whether the criticism is merited. Much of the criticism embodies a view that things used to be better years ago, when schools were not prey to the many defects they are alleged to show today. Recollections of the past may hide a mixed reality. In this article, criticisms of education from 1957 are compared with contemporary criticisms. Some issues have remained important across forty years, while a few new issues have emerged. Criticisms of forty years ago centered on the dominance of "professional educationists," progressivism, the life adjustment movement, the waning "spirit of competition," lax discipline, the lack of emphasis on classical and modern foreign languages, avoidance of science and math, the neglect of gifted children, the lack of training of children in moral and spiritual values, and low academic standards. Today's debates introduce the alleged test score declines, poor performance on international achievement comparisons, the supposed enormous increase in funding without positive results, the problem of high dropout rates, and the need to connect schooling and work. In addition, modern critics point to economic concerns, whether in terms of

funding for education or in regard to the contribution of schooling to economic development.

Educators are facing intense and elevated levels of criticism. Scholars debate whether the criticism is merited. The "pages" of the *Education Policy Analysis Archives* have themselves been a site for this debate on a number of occasions (e.g., Vol 1 No. 2; Vol 4 No 10, and the exchange between Lawrence Stedman and David Berliner in Volume 4, numbers 1, 3 and 7). Much of the criticism embodies a view that things used to be better years ago, when schools were not prey to the many defects they are alleged to show today. However, recollections of the past may be subject to a golden glow that hides a mixed reality.

A couple of years ago my colleague Hal May at The University of Manitoba retired after a long and productive career. Hal had amassed an impressive library of material in educational administration which he was busy trying to pass on to others--some to our Faculty library and other items to individuals he thought might make use of them. Among the pieces Hal gave me was a small pamphlet issued by the National Education Association in 1957. It was a Research Bulletin (Vol. XXXV, No. 4) from December, 1957, entitled "Ten Criticisms of Public Education." The authors were not identified, although the report lists Sam Lambert as director of research for the Association. The Research Bulletin was published three times per year for a subscription of $3 per year. (Prices are one thing that has changed since 1957!)

The NEA publication began with these words:

> Criticisms of public education in lay magazines and other publications have increased many fold in the past 10 years... . The initial step in counteracting destructive attacks and in utilizing valid criticisms for the improvement of public education is an objective appraisal of each criticism. (p.1)

The report authors selected ten published charges that appeared most frequently in a selection of 30 lay magazines from July 1954 to June 1957. For each of the ten, it gave a brief description of the charge and cited evidence and arguments to the contrary. Here are the ten charges that the NEA research group identified. Unlike the supposed list from the 1940s of 'ten worst problems of the schools' that Gerald Bracey debunked in the *Phi Delta Kappan* in 1994 (Bracey 1994), this list does have a documented source. Each statement of a criticism began with "some people say" or "critics say"; otherwise they are quoted exactly as written forty years ago.

1. ...the public schools are controlled or dominated by 'professional educationists' of schools of education, school superintendents, 'experts' in the state departments of education, 'specialists' in the US Office of Education, and the national organizations of

educationists.

2. ...John Dewey and 'progressive education' have taken over the public schools and that this philosophy of education is the chief cause of the crisis in education.
3. ...the life adjustment education movement is replacing intellectual training with soft social programs in most public- school systems.
4. ...the spirit of competition, an important incentive for learning, has been eliminated by the 100 percent annual promotion policy and the multiple-standard report card.
5. ...lax discipline in the public school is contributing to the increase of juvenile delinquency.
6. ... the teaching of classical and modern foreign languages is disappearing from the secondary schools.
7. ...high-school students, even the bright ones, are avoiding science and mathematics; fewer students are taking these courses now than 20 or 30 years ago.
8. ...the public schools are neglecting the gifted children because they are geared to teaching the average child.
9. ...the public schools are neglecting the training of children in moral and spiritual values.
10. ...the academic standards of schools of education are low: their programs of study are of questionable value, and the intellectual qualities of their students are the poorest in the universities.

The report went on to try to show how each of the charges was unjustified--not that the public schools were perfect, but that the issues were complex and the situation not nearly as bad as the critics were claiming. The report cited a range of data--for example, scores on college entrance tests--as well as studies by such luminaries as Lewis Terman, Robert Thorndike and James Bryant Conant. Many of the responses in the NEA report seemed similar to those being made today in reply to many of the same criticisms--that achievement levels were better than the media reported, that schools were trying to respond to extremely diverse student needs, that the serious problems were relatively few in number, that schools did stress moral values, that teacher education programs did attempt to impart appropriate skills, and so on.

Struck by the degree to which criticism of the schools seems unchanged from forty years ago, I decided to look more closely at how the criticisms and responses of 1957 actually did compare with those of 1997. To do this I reviewed some recent sources: the work of Diane Ravitch (1985, 1995; see Apple, 1996, for a review); Chester Finn (1991); E.D. Hirsch (1987); John Chubb and Terry Moe (1990). I also reviewed the work of some of the best known defenders of public schools, such as Gerald Bracey (1994, 1997), Richard Jaeger (1992) and David Berliner (in Berliner and Biddle, 1995) to see what they took to be the main criticisms being made. I made a similar review of Canadian sources--critics such as Mark Holmes (1992) and Andrew Nikiforuk (1993) as well as defenders such as Maude Barlow and Heather- Jane Robertson (1994). Finally, I examined a number of government and interest group reports on education in both countries, since these often embody criticisms of the current state of the schools.

Looking more carefully at the nature of the criticism of schools and the responses to it raised two issues. First, while some of the criticisms of schools being made today are very similar to those of 1957, others are new. The similarities and differences raise questions about the origins of criticism of schools. Second, reading criticism of schools over several decades raises questions about the nature of the debate over education, and especially the role of evidence. If the same issues surface again and again, does evidence matter? Does the debate itself matter? Is anybody listening? Without claiming to have any answers to these questions, I offer the following observations.

Criticism Then and Now

Certainly some of the criticisms on the 1957 list are still current. For example, reforms to governance in England and New Zealand were justified in part on the basis of excessive influence by professional educators; what is now called "provider capture." Teacher unions are often accused today of having undue influence and stifling reform because of self-interest. Another variant on this theme has to do with the supposedly baneful influence of school district bureaucracies, as argued by Chubb and Moe, or as a motive for the Chicago reforms of a few years ago. School-based management in some of its variants is also defended as moving authority to parents and/or teachers in the school, where real knowledge about problems and solutions is thought to reside.

Other points on the NEA list also continue to resonate in current debates. Progressivism remains a point of attack for many, such as Diane Ravitch (1985) in the U.S.A. or Mark Holmes (1992) in Canada, who argue that the move to child-centered education has resulted in lower standards. A number of critics--such as Diane Ravitch and E. D. Hirsch--are strong proponents of a purer academic mission for the school and less focus on "soft social programs". Despite strong evidence that retention in grade is ineffective, social promotion remains a controversial issue, and retention is still frequently supported by parents and teachers (Oakes, 1992). While the phrase "juvenile delinquency" has gone out of use, concerns about levels of violence in and around schools are high. *A Nation at Risk* cited low rates of enrolment in foreign languages, science and mathematics as very serious concerns, and concern about science and mathematics achievement has continued to be a prominent issue, taken up in many policy reviews and reform programs. Recent Canadian curriculum reforms have included greater time allocations for these subjects. Attention to the gifted also remains an issue. The U.S. Department of Education has recently issued reports dealing with this issue (Department of Education, 1996) and with the importance of science and mathematics (Department of Education, 1997). Finally, debate about teacher education continues, with many efforts in the U.S.A. on this front such as the work of the Holmes Group and efforts to create various state or national standards and licensing vehicles. Several provincial governments in Canada have also identified teacher education as a reform issue (though their proposals tend to be rather vague as to what the problems are). England also made dramatic changes in teacher education, moving much of the activity away from post-secondary institutions and placing it under the control of schools.

How Consistent Is Criticism of Schools?

In many respects, then, the issues of 1957 are also the issues of 1997, suggesting that criticism is eternal--and perhaps, by implication, not very meaningful. Those who think that the golden age was forty years ago (when they were young?) would surely be disappointed by the NEA's report. One suspects that the NEA authors little thought that their research would be as relevant in 1997 as it was four decades ago. have made the same point; Bracey and Berliner (in Berliner and Biddle, 1995) have both cited many earlier instances of criticism of schools going back to the early years of this century.

But that continuity is not the full story either. The debate in 1997 also includes some issues that were not on the agenda four decades ago. These include, most prominently, alleged test score declines, poor performance on international achievement comparisons, the supposed enormous increase in funding without positive results, the problem of high dropout rates, and the need for a stronger link between schooling and work. These issues feature prominently in current debates and are absent or muted in the 1957 NEA list.

The criticisms of the 1990s also have some very different preoccupations from those of the past, even when some of the specific manifestations are the same. Economic concerns, whether in terms of funding for education or in regard to the contribution of schooling to economic development, are central today and were much less so, it appears, forty years ago. Arguments for more language study or science education seemed then to be framed in terms of an image of the classically educated person; today they are framed in terms of the need for economic competitiveness.

The Argentinian writer Jorge Borges once wrote a story ("Pierre Menard") about a man in Argentina in the 1930s who had written a book that was word for word the same as *Don Quixote*. Borges wrote in the story that the two texts are "verbally identical but the second is almost infinitely richer." When Cervantes wrote such and such a phrase in the Spain of the sixteenth century, it had one meaning whereas when it was written in Buenos Aires in the 1930s it clearly carried associations from Neitzsche or William James or Bertrand Russell, who of course could not have influenced Cervantes! Borges was anticipating, perhaps, the postmodern view that a text takes on a new meaning when read in a new context. It does seem that concerns about such matters as the state of the gifted or the importance of values education can have quite a different significance in the current climate of economic insecurity and fear.

Diverse Critics

Lists of criticisms also distract us from the important observation that the critics are not all of one view. Some attack schools as being insufficiently traditional, while others regard them as insufficiently modern. Some think that schools should emphasize traditional academic pursuits while others seek a greater focus on specific workplace skills. Various commentators on the changes in education in England under Margaret Thatcher, for example, all note that the Conservatives themselves did not agree in their analyses of what was wrong with schools and what should be done to improve them (Lawton, 1994). Some were free-enterprisers who advocated market-based solutions while others were traditionalists who wanted a return to supposedly successful policies of an earlier era. Similarly, traditionalist critics such as E.D. Hirsch have quite a different analysis than do those focused on the economy, such as Marc Tucker or Willard Daggett. Are Christian conservatives really expressing a similar view to the National Governors' Association when each talks about the problems of standards in education?

Schools have also been subject to criticism from the left, or from non-conservative positions. For example, schools have been criticized for promoting or sustaining inequality, for failing to pay enough attention to diversity, and for inadequate concern for issues of social justice. A powerful example that affected my own early perceptions of schooling was work done in Toronto arguing that students from particular ethnic or economic backgrounds were being tracked into low-achievement programs.

In fact, many of the present defenders of schooling spent substantial earlier portions of their careers criticizing schools (Power, 1992). Many of the critics of neo-conservative education policies, who now defend schools as vital to maintaining equity, were themselves at one time highly critical of schools for failing to address social and economic inequities. In fact, there continues to be a vocal group of commentators who focus on the failure of schools to address problems of poverty and racism, although their voices are often lost in the much louder criticisms over standards and morals. Perhaps lists of criticisms actually hide a great deal of variability in the critics' analyses of schools. If so, such lists do not help us think clearly about the situation of and prospects for schools.

The Nature of Debate and the Role of Evidence

A second concern that grows out of an historical look at the debate over schooling has to do with the role of evidence in shaping our thinking.

The 1957 NEA report marshaled a considerable amount of empirical data in its attempt to refute charges against the schools. But the 1997 debate is much more evidence intensive. Not everyone relies on empirical evidence, of course. Where criticism of schools arises primarily from a religious or other value orientation, empirical data may play a much smaller role. But in reading the work of the critics and defenders of schooling, one cannot but be struck by the wide-ranging use of data. To take just one example, the debate about whether or not achievement levels in the United States have actually declined has featured many sophisticated competing analyses of several different data sources (e.g., Bracey, 1997; Stedman, 1996, 1997). Similarly, the argument about the impact of resources on achievement has involved a great deal of analysis of a large number of studies (Hanushek, 1994, 1997, also see the review by Gintis, 1995; Burtless, 1996).

All the evidence, however, does not seem to have resolved the arguments. In fact, more extensive evidence can have the effect of contributing to even more cynicism, as some people find the seemingly endless argument about the numbers reminiscent of the old saw that there exist prevaricators of three types: "liars, damn liars, and statisticians." Teachers and policy-makers may wonder whether research can ever inform policy and practice since even with much more evidence, the disagreement remains as heated as ever. Does research help? Does it matter? Does anyone really care about the data, since the same conclusions seem to be repeated by the same actors whatever evidence may be adduced? The role of evidence in resolving policy debates is a much-examined question (e.g., Anderson & Biddle, 1991: Stone, 1988). The consensus of current opinion would seem to be that evidence is only one factor in such matters--that issues of values and ideology are at least as important as evidence and may well shape what people are able or willing to see as evidence in the first place.

But we should not be too discouraged by this analysis. Empirical evidence may not answer all the questions for us, but it can help us answer some of them and think more deeply and more clearly about others. The ability of physically disabled students to learn in regular classrooms is no longer debated, largely because of conclusive evidence. In Canada, French Immersion programs (in which Anglophone students do almost all their schooling in French) were hotly debated in the 1970s and 1980s, but have been shown to be successful in developing academic skills in both languages.

Other issues are more controversial and thus less likely to be resolved by evidence. But here, too, over time evidence helps frame the debate. So, while debate continues about the value of tracking and streaming, the disproportionate placement of minorities in less challenging streams is agreed to be an aspect of the debate that requires attention. While arguments rage about comparative achievement levels, there is broad agreement that the simple tests of years ago are inadequate to assess the things that really matter. There may be no agreement on the importance of additional funding in promoting achievement, but there is growing acceptance that gross disparities in funding across schools and districts are undesirable. In almost every area of education policy, ideas have changed at least in part because of evidence.

Moreover, the search for evidence is itself a valuable activity, and one which reinforces the best ideals of education. Indeed, evidence-based arguments about policies are one of the main ways in which a society can learn, and so are especially important to encourage (Lindblom, 1990; Majone, 1989). The fact that critics and defenders of schools alike feel that it is vital to marshal evidence to support their position can be regarded as a step forward--a recognition that dogma is not enough and that there is at least the possibility of subjecting disparate ideas to a test whose legitimacy is widely upheld. For much of human history, disputes in viewpoint have been resolved through isolation or through violence--we avoid or conquer those who disagree. Debate, even if it is acrimonious, can be seen as an important human achievement. In another context, Nicholas Burbules and Suzanne Rice (1991) have written eloquently about the importance of "dialogue across difference." Nobody should be surprised that fundamental differences in values are not quickly resolved through evidence, but we should all be pleased that evidence is seen to be important and at least in principle accepted as a basis for bridging differences.

Conclusion

Schools will always be the subject of intense criticism for at least two reasons. First, society's goals for schools are extremely ambitious. In an important sense, schooling is about perfection. We hope that our schools can do everything--shape young people who are thoughtful, productive, articulate, considerate, knowledgeable, patriotic, worldly, idealistic, realistic, challenging, accepting, critical, loyal. We expect our schools to teach our children knowledge, skills, and values, but also to overcome the same social problems that we as adults have been unable to solve--to reduce poverty, to build the economy, to save the environment, to include the excluded, to look after oneself and care for others, to overcome materialism.

Second, people do not agree about which of these goals are most important or about how any given goal is best accomplished. Some want to stress individual excellence and others, social equity. Some emphasize traditional academic learning and others want to focus on the emerging needs of the economy. Some may value most patriotism and loyalty while others give priority to independent thought and critical thinking. It is no accident that most lists of goals for schools contain a large number of items that are not always mutually consistent. And the growing diversity in our society, coupled with the growing recognition of the importance of diversity, makes the challenge steadily greater as we struggle to develop a common institution that is also able to accommodate difference (Levin & Riffel, 1994; Riffel, Levin & Young, 1996).

Schools cannot achieve all the things we want from them, and they cannot satisfy all the expectations we have of them. They will inevitably be the objects of criticism. And the more important our goals for schools are, the more intense the criticism is likely to be. The paradox here is that criticism is actually a sign of respect. I've had occasion to remind school administrators that the increased willingness of parents to challenge school policies and practices is an indicatioin of the success of education. After all, we hope that schools will help people learn to define, articulate and work for what they hold to be important. We should be pleased when they do so, even if it makes our lives harder. If people thought schools unimportant, they would not take the time to argue about their achievements and shortcomings.

Rather than frustration and despair over criticism, then, we might benefit from seeing criticism as an opportunity--a chance to create discussion about things that are important, to help us achieve the vital educational task of learning to live together even with all our differences. Certainly criticisms can be unfair, mischievous, or even malevolent. Defenders of public schools should continue to speak out and to bring to bear arguments and evidence in support of their views. But we will all benefit insofar as we can see debate as having the potential to move us in a desirable direction.

What do we learn, then, from looking at criticisms of education today and forty years ago? We learn that some issues remain important and new issues emerge. We learn that our ability to define, understand and debate issues is imperfect. We learn that the schools probably face an impossible task. But we also learn that people care about education, that evidence and reason can make a difference, and that the struggle for better education remains a vital enterprise.

References

Anderson, D. and Biddle, B. (Eds) (1991). *Knowledge for Policy: Improving Education Through Research*. London: Falmer Press.

Apple, M.W. (1996). Being Popular About National Standards: A Review Of Ravitch's *National Standards in American Education: A Citizen's Guide. Educational Policy Analysis Archives* Vol. 4 No. 10. (Available online: http://epaa.asu.edu/epaa/v4n10.html).

Barlow, M. & Robertson, H-J. (1994). *Class warfare: The assault on Canada's schools*. Toronto: Key Porter.

Berliner, D.C. and Biddle, B.J. (1995). *The Manufactured Crisis: Myths, Fraud, and the Attack on America's Public Schools*. Reading, MA: Addison-Wesley.

Berliner, D.C., and Biddle, B.J. (1996). Making molehills out of molehills: Reply to Lawrence Stedman's review of *The manufactured crisis. Educational Policy Analysis Archives* 4 (February 26). (Available online: http://epaa.asu.edu/epaa/v4n3.html).

Bracey, G.W. (1991). Why can't they be like we were? *Phi Delta Kappan* 73: 104-17.

Bracey, G.W. (1992). The second Bracey report on the condition of public education. *Phi Delta Kappan* 74: 104-8, 110-17.

Bracey, G.W. (1993). The third Bracey report on the condition of public education. *Phi Delta Kappan* 75: 104-12, 114-18.

Bracey, G.W. (1994). The fourth Bracey report on the condition of public education. *Phi Delta Kappan* 76: 115-27.

Bracey, G.W. (1995). The fifth Bracey report on the condition of public education. *Phi Delta Kappan* 77: 149-60.

Bracey, G.W. 1996. The sixth Bracey report on the condition of public education. *Phi Delta Kappan* 78: 127-38.

Bracey, G.W. (1997a). The seventh Bracey report on the condition of public education. *Phi Delta Kappan* 79: 120-36.

Bracey, G.W. (1997b). Rejoinder: Comparing the incomparable: a response to Baker and Stedman. *Educational Researcher*, 26(3), 19-26.

Burbules, N. and Rice, S. (1991). Dialogue across differences: Continuing the conversation. *Harvard Educational Review*, 61(4), 393-416.

Burtless, G. (Ed) (1996) Does money matter? The effect of school resources on student achievement and adult success. Washington: Brookings Institute.

Chubb, J. E. & Moe, T. M. (1990). *Politics, markets, and America's schools.* Washington, DC: The Brookings Institution.

Finn, C.E. (1991). *We must take charge: Our schools and our future.* New York: Free Press.

Gintis, H. (1995). Review of Eric A. Hanushek's *Making Schools Work. Educational Policy Analysis Archives* Vol. 3 No. 7. (Available online: http://epaa.asu.edu/epaa/v3n7.html).

Hanushek, E.A. (1994). *Making Schools Work: Improving*

Performance and Controlling Costs. Washington, DC: The Brookings Institutions.

Hanushek, E.A. (1997). Assessing the effects of school resources on student performance: An update, *Educational Evaluation and Policy Analysis*, 19(2), 141-164.

Hirsch, E.D. (1987). *Cultural literacy: What every American needs to know.* Boston, MA: Houghton Mifflin.

Holmes, M. (1992). The revival of school administration: Alasdair MacIntyre in the aftermath of the common school. *Canadian Journal of Education*, 17(4), 422-436.

Jaeger, R.M. (1992). World class standards, choice, and privitization: Weak measurement serving presumptive policy. *Phi Delta Kappan*, (October), 118-128.

Lawton, D. (1994). *The Tory Mind on Education 1979-1994.* London: Falmer Press.

Levin, B. and Riffel, J.A. (1994). Dealing with diversity: Some propositions from Canadian education *Education Policy Analysis Archives.* Vol. 2, No. 2. (Available online at http://epaa.asu.edu/epaa/v2n2.html).

Lindblom, C. (1990). *Inquiry and Change.* New Haven: Yale University Press .

Majone, G. (1989). *Evidence, Argument and Persuasion in the Policy Process.* New Haven: Yale University Press.

Nikiforuk, A. (1993). *School's out: The catastrophe in public education and what we can do about it.* Toronto: McFarlane Walter Ross.

Oakes, J. (1992) Can tracking research inform practice? Technical, normative, and political considerations. *Educational Researcher.* 21(4), 12-21.

Power, S. (1992) Researching the impact of education policy: difficulties and discontinuities. *Journal of Education Policy*, 7(4), 493-500

Ravitch, D. (1985). *The Schools We Deserve* (New York: Basic Books, 1985).

Ravitch, D. (1995) *National Standards in American Education: A Citizen's Guide.* Washington: The Brookings Institution.

Riffel, J., Levin, B. and Young, J. (1996). Diversity in Canadian education. *Journal of Education Policy*, 11(2), 113- 123.

Stedman, L.C. (1996a). The Achievement Crisis is Real: A Review of *The Manufactured Crisis*. *Education Policy Analysis Archives*. Vol. 4, No. 1. (Available online at http://epaa.asu.edu/epaa/v4n1.html).

Stedman. L.C. (1996b). Respecting the Evidence: The Achievment Crisis Remains Real. *Education Policy Analysis Archives*. Vol. 4, No. 1. (Available online at http://epaa.asu.edu/epaa/v4n7.html).

Stedman, L. (1997). International achievement differences: An assessment of a new perspective. *Educational Researcher*, 26(3), 4-15.

Stone, D. (1988). *Policy Paradox and Political Reason*. New York: HarperCollins.

U. S. Department of Education (1996). National Excellence: A Case for Developing America's Talent. Washington: U.S. Department of Education.

U. S. Department of Education (1997). The Condition of Education 1997. Washington: U.S. Departrment of Education

About the Author

Benjamin Levin
The University of Manitoba

Email: levin@electra.cc.umanitoba.ca

Benjamin Levin is professor of educational administration and Dean of Continuing Education at The University of Manitoba. His main academic interests are in education politics, policy and economics. He is a long-time participant on the Education Policy Analysis Archives editorial board.

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

John Covaleskie
Northern Michigan University

Alan Davis
University of Colorado, Denver

Mark E. Fetler
California Commission on Teacher Credentialing

Thomas F. Green
Syracuse University

Arlen Gullickson
Western Michigan University

Aimee Howley
Marshall University

William Hunter
University of Calgary

Daniel Kallós
Umeå University

Thomas Mauhs-Pugh
Rocky Mountain College

William McInerney
Purdue University

Les McLean
University of Toronto

Anne L. Pemberton
apembert@pen.k12.va.us

Richard C. Richardson
Arizona State University

Dennis Sayers
Ann Leavenworth Center
for Accelerated Learning

Michael Scriven
scriven@aol.com

Robert Stonehill
U.S. Department of Education

Greg Camilli
Rutgers University

Andrew Coulson
a_coulson@msn.com

Sherman Dorn
University of South Florida

Richard Garlikov
hmwkhelp@scott.net

Alison I. Griffith
York University

Ernest R. House
University of Colorado

Craig B. Howley
Appalachia Educational Laboratory

Richard M. Jaeger
University of North Carolina--Greensboro

Benjamin Levin
University of Manitoba

Dewayne Matthews
Western Interstate Commission for High
Education

Mary McKeown-Moak
MGT of America (Austin, TX)

Susan Bobbitt Nolen
University of Washington

Hugh G. Petrie
SUNY Buffalo

Anthony G. Rud Jr.
Purdue University

Jay D. Scribner
University of Texas at Austin

Robert E. Stake
University of Illinois--UC

Robert T. Stout
Arizona State University

---

archives | abstracts | editors | board | submit | comment | subscribe | search

This article has been retrieved 1080 times since September 6, 1998.

Education Policy Analysis Archives

Volume 6 Number 17 **September 6, 1998** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal.
Editor: Gene V Glass Glass@ASU.EDU.
College of Education
Arizona State University,Tempe AZ
85287-2411

Copyright 1998, the EDUCATION POLICY ANALYSIS ARCHIVES.Permission is hereby granted to copy any article provided that EDUCATION POLICY ANALYSIS ARCHIVES is credited and copies are not sold.

Performance Indicators: Information in Search of a Valid and Reliable Use

E. Raymond Hackett
Auburn University

Sarah D. Carrigan
University of North Carolina at Greensboro

Abstract Measures of overall institutional performance were explored from a decision support perspective with twenty similar Carnegie Classification Baccalaureate II institutions. The study examined the usefulness of performance indicators in campus decision making following both a hypothesis testing and case study approach. Two conclusions were reached: first, that the performance measures most commonly cited in the literature as measures of institutional financial viability are of limited use for institution specific policy development; and second, that performance indicators are most effectively used within an institution specific, whole system framework.

Introduction

State-defined performance indicators for institutions of postsecondary education are rapidly becoming the hallmark of the 1990s. By 1993 over one-third of the states had some form of performance indicator legislation enacted (Bogue, Creech & Folger, 1993) and with each legislative session since the number has increased. Discussion at the state level has begun to shift toward funding the enterprise based on outcomes, effectiveness, and efficiency (Gather, Nedwek, and Neal, 1994). Significant attempts at operationalizing these concepts and weaving them into the fabric of planning, policy and budget development were given license in several states during the 1997 and 1998 legislative sessions.

In the 1994 Education Commission of the States publication Charting Higher Education Accountability (Ruppert, 1994) a case study of ten states indicated that the adoption of state-level performance indicators most often was done rapidly, relied on existing data and usually was driven by legislative initiative. This report implied that few states have accomplished the analysis necessary to define measures appropriate for systemic decision making and public reporting.

With the advent of student right-to-know legislation, federally defined performance indicators for institutions of postsecondary education became a larger part of the institutional reporting cycle. In 1996, the Department of Education proposed a far more explicit use of performance indicators, and this proposal led to a national debate. The belief that a unique equation could provide an indication of institutional financial and programmatic health, and that institutional scores on a specific set of indicators should impact the disbursement of federal funds, was outlined in the Federal Register Volume 61, Number 184 on September 20, 1996. In this Notice of Proposed Rulemaking, the Secretary of Education proposed to amend the Student Assistance General Provision regulations by revising the requirements for compliance audits and adding a new subpart establishing financial responsibility standards. The proposed regulations would require institutions participating in programs authorized by Title IV of the Higher Education Act of 1965, as amended, to meet cutoff scores on certain calculated financial ratios to avoid a compliance audit.

Certainly, institutions of postsecondary education should be held accountable to their constituents, their service area, and the public that provides monetary and other support. However, there is a concomitant reality, that is the reality of the deans, administrators, faculty and staff attempting to manage real institutions in a real world. At this level there is only one question. How do I make good decisions? And that is a powerful question. For it is the sum of the decisions made during the campus year that create the future for an institution. It is the sum of these decisions that lead to outcomes, effectiveness, and efficiency. It is at the decision point where institutional research finds its home and performance indicators have meaning. Offices of institutional research conduct studies and convert data into information for two primary purposes: to support the decision making process by providing analyses that serve to reduce uncertainty prior to making a decision; and to assess how effective the institution has been at meeting the goals and objectives outlined in the campus plan. The former is for internal constituents and the latter for both internal and external constituents.

While there is significant research in postsecondary education on the development of information to support an understanding of the operation and outcomes of the enterprise, further research must be focused on defining decision points. With a taxonomy of decision points, and an understanding of how they interrelate, research can be focused on the amount of uncertainty that is reduced by various performance indicators at given decision points. A clearer understanding of performance indicators and their relationship to decision support must be developed.

This article approaches the use of performance indicators from two perspectives. In the first study eleven frequently cited performance indicators were used to explore the implications of enrollment stability and financial viability with twenty similar Carnegie Classification Baccalaureate II institutions. This study examined issues addressed in the Federal Register Volume 61 proposal to amend the Student Assistance General Provision regulations by revising the requirements for compliance audits and adding a new subpart establishing financial responsibility standards. The implication here was that institutional scores on a specific set of indicators define the financial viability of an institution and should impact the disbursement of federal funds.

The second study used a case study approach to focus on a campus included in the sample of institutions used in the first study. In this particular case study, the institution had decided that challenges on two fronts were threatening the institution. The institution moved to change both the population of students served and the focus of the academic program. The use of information and performance indicators to support decisions related to the repositioning was explored.

Study One

Review of the Problem and Literature

Measures of academic programs, staffing, enrollment level, student and faculty characteristics, and revenue and expense can help define an institution's programmatic and financial strengths and weaknesses. At independent institutions, particularly the smaller liberal arts institutions, it is essential that the campus leadership understand the implications of these numeric indicators and their interrelationships. A significant change in the value of key performance indicators at smaller institutions can signify changes that will impact the campus for a given year, or a number of years. With the publishing of the National Association of College and University Business Officers (NACUBO), Financial Self- Assessment: A Workbook for Colleges and Universities, in the early 1980s a move began to understand the campus and campus policy in terms of performance indicators. Certainly, the total quality improvement concept of benchmarks falls along the continuum of work that has been conducted on performance indicators.

There has been significant discussion on the development of performance indicators and their use in higher education. Among the extant models are: the National Association of College and University's Financial Self-Assessment Workbook (1987); Performance Measurement Systems for Higher Education (Kidwell and Long, 1995); Strategic Indicators for Higher Education (Taylor, Myerson and Massy, 1993); and Measuring Up: The Promises and Pitfalls of Performance Indicators (Gather, Nedwek and Neal, 1994). The Joint Commission on Accountability Reporting (JCAR), a project of the American Association of State Colleges and Universities, the American Association of Community Colleges, and the National Association of State Universities and Land-Grant Colleges has produced a framework for accountability reporting recently summarized in the 1996 publication JCAR Technical Conventions Manual. Currently in progress is the NACUBO Benchmarking Project, which is developing quantitative measures to set as a point of reference and standard for basic operations. However, most of the analysis and literature on the development of state defined institution-level performance indicators describes a pattern of implementation with little prior conceptual development and a focus on interinstitutional comparison (Bogue, Creech & Folger, 1993). A 1994 Education Commission of the States study found that performance indicator initiatives in the various states contain many of the same measures (Ruppert, 1994). Most of the states studied used 20 or so indicators that were collected by a governing board and reported in a tabular form. The indicators most commonly used reflected some measure of: instructional inputs; instructional process and use of resources; instructional outcomes; efficiency and productivity; diversity and access; articulation; and relation to state needs.

In the 1987 revision of Financial Self- Assessment: A Workbook for Colleges and Universities (Dickmeyer & Hughes), the concept of an overall institutional equation, defined in terms of key performance indicators, was again emphasized. It was strongly implied in this volume that there were ranges within the various indicators presented that indicated good, moderate or poor performance on a given indicator. It was also implied, in this major work of a standing NACUBO committee, that a certain equation could be inferred for an institution from a combination of these indicators. It was further implied that this unique equation could provide an indication of institutional health, and areas of institutional strength and weakness. Since 1987 a number of institutions have adopted the self-assessment strategy put forth in this volume and a modest research literature has developed. A noticeable addition to this strategy was put forth by Mary Sapp (1994) in the AIR Professional File document, Setting a Key Success Index Report: A How to Manual. A recasting of standard financial ratios to accommodate the Financial Accounting Standards Board's Statement of Financial Accounting Standards No. 116 and 117 was accomplished by Prager, McCarthy and Sealy (1995). These new ratios were cited in the proposal to amend the Student Assistance

General Provision regulations.

Given the intense interest in performance indicators, there is a surprising lack of literature examining the policy relevance of indicators of institutional performance. There are few articles assessing the construct validity of the widely used Dickmeyer and Hughes (1987) NACUBO publication. The first study explores a portion of the financial viability framework outlined in Financial Self-Assessment: A Workbook for Colleges and Universities using 20 Baccalaureate II institutions that are members of the College Information Systems Association.

The College Information Systems Association is an association of over 30 liberal arts institutions, each with less than 2,500 FTE students, that share a common data set and share a research office staffed by faculty and graduate students in a higher education doctoral program at a land-grant university. These institutions provide an excellent laboratory for testing performance indicators. These liberal arts colleges and universities are similar in size, age of institution, and academic program offerings. Financial and enrollment data from these institutions offer few confounds. The students are overwhelmingly full-time and seeking the bachelors degree. Revenues come in the form of tuition and fees, gifts, endowment earnings, and auxiliary enterprise charges. Expenses are primarily for faculty and staff, academic support, student life, and physical plant. The primary units of production are the student credit hour and headcount student.

The purpose of this study was to examine the concept of key performance indicators with institutional viability defined in terms of enrollment stability and the ability to meet financial obligations.

Methods

Data were collected from twenty institutions of the College Information Systems Association for a five year period from FY 1992-93 through FY 1996-97 and included 265 measures. Most of these measures were data already being supplied by the colleges to National Center for Educational Statistics and other national organizations such as the Council for the Advancement and Support of Education. Data were collected on revenues, private support, expenditures and transfers, balance sheet items, plant, personnel, faculty development, instruction, student characteristics, financial aid, library holdings, and data processing equipment. A data element dictionary recapitulating and refining national definitions was prepared and taught to the institutions through a series of workshops. Institution level performance indicators were developed from primary data and took the form of primary data; totals of primary data; percentages of total; ratios; and appropriate algorithmic transformations. At the time of the study only twenty of a possible 26 had reported and verified data for the five years under study. The association staff have discovered through this project the difficulty of collecting accurate, timely and comparable data from a number of institutions even with national data standards. Each campus has a number of primary data providers and that will confound any study over multiple campuses.

For the purposes of this preliminary investigation, performance indicators carry the maximum of information when they provide the decision making process with insight into whether an institution is maintaining a steady level of viability; losing viability; or gaining viability. Institutions were defined as viable if they maintain enrollment and maintain financial viability. Of course an essential element of institutional viability is whether an institution is meeting the goals and measurable objectives outlined in the campus plan. Assessing institutional outcomes in terms of consistency with institutional goals was outside the scope of this study.

The first step in this investigation was to develop a set of core indicators that could provide an indication of institutional viability. In Measuring Up: The Promises and Pitfalls of Performance Indicators (Gaither, Nedwek and Neal, 1994) ten core indicators that are found in use and cited most frequently as measures of institutional viability are listed. In the present study that list was modified slightly to focus on the institutional viability construct outlined in Dickmeyer and Hughes (1987). This construct focused on enrollment stability and flexibility in managing available revenues, funds, and expenditures. An eleventh indicator, percent change in fall fulltime equivalent students (fte), was included with the core indicators in order to explore the concept of institutional viability as defined in this study. The eleven indicators used are defined in Table 1.

Table 1
Definition of the eleven performance indicators used in study one.

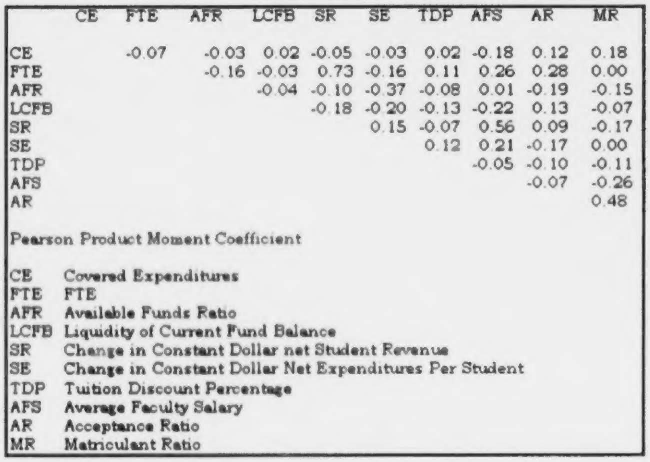
1. Covered Expenditures - Excess (deficit) of current fund revenues over (under) current fund. expenditures
2. FTE - Fall full-time equivalent students.
3. Percent Change FTE - Percent change in fall full-time equivalent students over previous year
4. Constant Dollar Net Student Revenue - Total tuition and fee revenues minus unrestricted current fund scholarships and fellowships adjusted by the HEPI.
5. Constant Dollar Net Expenditures per Student - Total current fund expenditures and transfers adjusted by the HEPI index divided by fall FTE.
6. Tuition Discount Percentage - Defined as: ((tuition & fee revenues minus unrestricted current fund scholarships an & fellowships) divided by full-time tuition and fee rate)) divided by fall FTE students.
7. Available Funds Ratio - Defined as: (sum of the unrestricted current fund balance, quasi-endowment at market value, and unexpended plant fund balance) divided by unrestricted education and general expenditures plus mandatory transfers.
8. Liquidity of the Current Fund Balance - Defined as: cash in the unrestricted current fund plus investments in the unrestricted current fund divided by liabilities in the unrestricted current fund.
9. Average Faculty Salary - Average salary for all full-time faculty.
10. Acceptance Ratio - Number accepted divided by number applied.
11. Matriculant Ratio - Number matriculated divided by number accepted.

The identified indicators were first examined using descriptive statistics and analysis of variance across all the institutions for five years. The institutions were then divided into two groups defined in terms of their viability based on the stability of the student population and the institution's financial position. For the student population, stability was defined in terms of number of enrolled students and change in number of enrolled students. Financial viability was defined in terms of the institution's ability to meet its financial obligations without significantly changing fund balances and by the NACUBO ratio level definition for liquidity of the current fund balance and availability of fund balances to meet current obligations. The performance indicators defined were then compared within the new groupings of institutions and financial viability and enrollment stability examined using descriptive statistics and multivariate statistics.

Results

The twenty institutions for which valid and reliable data were available were all Carnegie Classification Baccalaureate II and similar in academic program offerings. A Pearson Product Moment Correlation Coefficient was calculated for each of the chosen core indicators in relationship with each other for all the institutions for all years. That matrix, found in Figure 1, indicated only four relationships of any magnitude: (1) enrollment was positively related to net student revenues; (2) average faculty salary was positively related to net student revenues; (3) the matriculant ratio was positively related to the applicant ratio;

(4) the available funds ratio was negatively related to expenditures per student.



	CE	FTE	AFR	LCFB	SR	SE	TDP	AFS	AR	MR
CE		-0.07	-0.03	0.02	-0.05	-0.03	0.02	-0.18	0.12	0.18
FTE			-0.16	-0.03	0.73	-0.16	0.11	0.26	0.28	0.00
AFR				-0.04	-0.10	-0.37	-0.08	0.01	-0.19	-0.15
LCFB					-0.18	-0.20	-0.13	-0.22	0.13	-0.07
SR						0.15	-0.07	0.56	0.09	-0.17
SE							0.12	0.21	-0.17	0.00
TDP								-0.05	-0.10	-0.11
AFS									-0.07	-0.26
AR										0.48

Pearson Product Moment Coefficient

CE	Covered Expenditures
FTE	FTE
AFR	Available Funds Ratio
LCFB	Liquidity of Current Fund Balance
SR	Change in Constant Dollar net Student Revenue
SE	Change in Constant Dollar Net Expenditures Per Student
TDP	Tuition Discount Percentage
AFS	Average Faculty Salary
AR	Acceptance Ratio
MR	Matriculant Ratio

Figure 1. Pearson product moment correlation coefficients for the ten core indicators.

The first three of these relationships might have been expected. The implication that institutions with a stronger available funds position were expending less per student, though understandable, certainly warranted further study. The lack of other relationships was considered the strongest indication that further study was warranted.

In terms of an overall profile, Figure 2 details five years of percent change in fall fte data for the institutions. Figure 3 details five years of covered expenditures, or the excess (deficit) of current fund revenues over (under) current fund expenditures. As can be seen in Figure 2, in almost every case the institutions managed to maintain or expand enrollment over the five year period. The financial data presented in Figure 3 suggests that two distinct groups could be developed based on ability to meet expenditure demands with available revenues.

Figure 2. Percent change in fall FTE from previous year.

BEST COPY AVAILABLE

	FY 92-93	FY 93-94	FY 94-95	FY 95-96	FY 96-97
College A	0.14%	6.60%	9.35%	0.00%	4.02%
College B	1.22%	6.65%	5.29%	-9.87%	0.82 %
College C	-0.98%	-1.23%	2.75%	4.87%	1.3 5%
College D	7.66%	-3.20%	4.50%	2.82%	2.95 %
College E	-20.21%	-8.39%	37.94%	2.45%	2.95%
College F	18.63%	3.14%	1.52%	-3.75%	4.89%
College G	-8.47%	-9.72%	-0.31%	-0.56%	- 4.77%
College H	-6.79%	0.13%	6.60%	3.03%	0.75 %
College I	-0.36%	2.61%	-7.18%	-3.40%	-2 .08%
College J	0.95%	-0.42%	1.29%	117.53%	29.84%
College K	9.58%	1.73%	-10.88%	8.94%	2.34%
College L	N/A	N/A	N/A	N/A	N/A
College M	6.54%	-0.60%	0.35%	.62%	1.98%
College N	N/A	N/A	N/A	N/A	N/A
College O	11.07%	10.08%	13.73%	-1.25%	8.41%
College P	26.77%	6.99%	8.56%	16.18%	14.62%
College Q	5.04%	-2.40%	-2.69%	-2.40%	-0 .61%
College R	3.65%	-3.44%	-3.57%	3.82%	0.1 1%
College S	-0.59%	4.30%	-1.69%	-2.17%	-0 .04%
College T	-3.14%	-6.78%	3.38%	-9.59%	-4 .04%

Figure 3. Covered expenditures, FY 1992-93 to FY 1996-97.

	FY 92-93	FY 93-94	FY 94-95	FY 95-96	FY 96-97
College A	$24,071	($222,895)	$322,656	($305,060)	$1,967,207
College B	$487,342	$931,612	$1,075,006	($93,044)	$2,421,350
College C	($191,674)	($83,843)	$5,298	$85,799	$1,019,395
College D	$15,881	($295,349)	($109,014)	($301,853)	($22,474)
College E	($180,268)	($955,771)	$118,237	$135,813	$1,105,218
College F	($376,915)	$1,185,143	$128,121	($2,205)	($118,892)
College G	($408,189)	$1,654,871	($359,433)	($1,263,470)	($689,241)
College H	$1,159,043	$28,428	$1,009,956	$619,496	$678,227
College I	$49,512	$10,728	($300,306)	$12,480	($456,712)
College J	$1,570,039	$2,090,371	$1,578,264	$1,533,880	$210,909
College K	$217	$2,707	$1,068	$3,692	$102,458
College L	$808,590	$770,227	$361,301	$452,010	($398,757)
College M	$753,339	$372,701	$198,771	$468,484	($179,087)
College N	($1,065,616)	($202,832)	($1,861,550)	($1,699,522)	($1,173,749)
College O	$434,207	$707,564	$300,052	$1,772,279	$951,086
College P	$265,177	$154,010	$136,554	$264,129	($74,804)
College Q	($56,006)	($325,000)	($68,577)	$33,890	($115,278)
College R	($291,984)	($389,348)	($343,438)	($994,877)	($1,892,805)
College S	$221,497	$212,235	$799,562	$878,162	$245,226
College T	$66,847	($395,514)	($783,982)	$5,523	$68,381

After reviewing the Covered Expenditure data together with the available funds ratio and liquidity of the current fund balance, the institutions were divided into two groups. One group was designated as the strong group and consisted of ten institutions that were able to consistently maintain financial viability as indicated by covered expenditures, available funds ratio and liquidity of the current fund balance. The second group was designated the weak group and consisted of ten institutions that were not able to consistently maintain financial viability as indicated by covered expenditures, available funds ratio and liquidity of the current fund balance. These two groups were used to explore the relationship between the construct institutional viability, defined in terms of the three financial measures and fte, and the information provided by the selected performance indicators.

It was decided to use multiple linear regression to begin to define sets of information that might be related to institutional viability using the two groups identified. The three financial measures and fte were used as dependent variables and each of the ten indicators compared individually as independent variables for all institutions in each group, for all five years. Independent variables with a significant R^2 (F-test) and P-value were placed in a multiple linear equation as independent variables with the related dependent variables. The eight independent variables were:

1. Covered expenditures - strong group.
2. Covered expenditures - weak group.
3. Liquidity of the current fund balance - strong group.
4. Liquidity of the current fund balance - weak group
5. Available funds ratio - strong group
6. Available funds ratio - weak group
7. Full-time equivalent students - strong group
8. Full-time equivalent students - weak group

As can be seen in Figure 4, for the institutions that were able to consistently maintain financial viability the dependent variable, covered expenditures, was positively related with the matriculant ratio. This would indicate that the size of the freshman class over the five year period had a significant though small (adjusted $R^2 = .164$) impact on a balanced budget. For the institutions that were not able to consistently maintain financial viability the dependent variable, covered expenditures, was only positively related with the acceptance ratio. The implication here is that becoming less selective over the five year period had a significant though inconsequential (adjusted $R^2 = .066$) impact on decreasing budgetary imbalances. What was interesting in this analysis was not only the minimal impact of the noted effects, but also that none of the other independent variables had an effect for this dependent variable for either group.

Figure 4. Significant results for the dependent variable: covered expenditures.

Strong group.

Regression Statistics	
Multiple R	0.432
R Square	0.186
Adjusted R Square	0.164
Standard Error	591099.732
Observations	39

ANOVA

	df	SS	MS	F	Significance F
Regression	1	2.963E+12	2.963E+12	8.481E+00	6.050E-03
Residual	37	1.293E+13	3.494E+11		
Total	38	1.589E+13			

	Coefficients	Standard Error	t Stat	P-value
Intercept	-455867.482	376535.084	-1.211	0.234
Matriculant Ratio	2092804.983	718621.433	2.912	0.006

Weak group.

Regression Statistics	
Multiple R	0.298
R Square	0.089
Adjusted R Square	0.067
Standard Error	605017.940
Observations	44

ANOVA

	df	SS	MS	F	Significance F
Regression	1	1.496E+12	1.496E+12	4.087E+00	4.961E-02
Residual	42	1.537E+13	3.660E+11		
Total	43	1.687E+13			

	Coefficients	Standard Error	t Stat	P-value
Intercept	678212.270	399274.801	1.699	0.097
Acceptance Ratio	-2333811.738	1154382.406	-2.022	0.050

For the institutions that were able to consistently maintain financial viability the dependent variable, liquidity of the current fund balance, was positively related to the three independent variables, percent change fte, constant dollar net expenditures per student, and average faculty salary (adjusted $R^2 = .288$), as seen in Figure 5. This would indicate that a consistent growth in the size of the student population is related to financial strength in these institutions. For the institutions that were not able to consistently maintain financial viability, the liquidity of the current fund balance was positively related to the independent variables, covered expenditures and constant dollar net student revenue. The positive relationship evidenced by these two financial variables would be expected. What is interesting is the modest amount of variance that is accounted for (adjusted $R^2 = .196$) by two financial variables that should have a strong relationship with this measure of institutional viability. This could be construed as a fairly explicit indication that other expenditure related pressures must be considered in reviewing the financial viability of these institutions, institutions that have been unable to balance revenue to expense on a consistent basis. As with covered expenditures, what was interesting in this analysis was not only the modest impact of the noted effects, but also that none of the other independent variables had an effect for this dependent variable for either group.

Figure 5. Significant results for the dependent variable: liquidity of the current fund balance.

Strong group.

Regression Statistics	
Multiple R	0.598
R Square	0.357
Adjusted R Square	0.288
Standard Error	5.834
Observations	32

ANOVA	df	SS	MS	F	Significance F
Regression	3	529.239	176.413	5.183	0.006
Residual	28	953.016	34.036		
Total	31	1482.255			

	Coefficients	Standard Error	t Stat	P-value
Intercept	13.687	8.186	1.672	0.106
Percent Change in FTE	38.001	15.238	2.494	0.019
Expenditures per Student	-0.001	0.001	-1.517	0.141
Faculty Salary	0.000	0.000	-0.716	0.480

Weak group.

Regression Statistics	
Multiple R	0.479
R Square	0.230
Adjusted R Square	0.197
Standard Error	1.417
Observations	50

ANOVA

	df	SS	MS	F	Significance F
Regression	2	28.131	14.065	7.007	0.002
Residual	47	94.347	2.007		
Total	49	122.478			

	Coefficients	Standard Error	t Stat	P-value
Intercept	1.789	0.478	3.740	0.000
Covered Expenditures	0.000	0.000	-2.866	0.006
Student Revenues	0.000	0.000	-1.861	0.069

As can be seen in Figure 6 below, for the institutions that were able to consistently maintain financial viability the dependent variable, available funds ratio, was positively related with full- time equivalent students and tuition discount percentage. This seems to imply that size of the student population, maintained by leveraging tuition, is related to overall institutional financial strength. This effect was one of the larger effects seen in this study (adjusted $R^2 = .349$). For the institutions that were not able to consistently maintain financial viability the dependent variable constant dollar net expenditures per Student was the only independent variable positively related with the available funds ratio (adjusted $R^2 = .197$).

Figure 6. Significant results for the dependent variable: available funds ratio.

Strong group.

Regression Statistics	
Multiple R	0.614
R Square	0.376
Adjusted R Square	0.349
Standard Error	1.686
Observations	49.000

ANOVA					
	df	SS	MS	F	Significance F
Regression	2	78.886	39.443	13.884	0.000
Residual	46	130.682	2.841		
Total	48	209.568			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%
Intercept	4.513	0.682	6.623	0.000	3.142
FTE	-0.002	0.000	-3.582	0.001	-0.002
TDP	-5.120	1.236	-4.143	0.000	-7.607

Weak group.

Regression Statistics	
Multiple R	0.439
R Square	0.193
Adjusted R Square	0.173
Standard Error	72.699
Observations	42

ANOVA

	df	SS	MS	F	Significance F
Regression	1	50506.737	50506.737	9.556	0.004
Residual	40	211405.735	5285.143		
Total	41	261912.472			

	Coefficients	Standard Error	t Stat	P-value
Intercept	81.084	24.070	3.369	0.002
Expenditures per Student	-0.014	0.004	-3.091	0.004

The implication of the above results is that, within this group, those institutions that have a higher expenditure level are also financially more viable. The available funds ratio is perhaps the single best measure of an institutions financial viability in that it accounts for all funds that could be marshaled to meet institutional financial obligations. What stands out, using this most inclusive of financial measures, is that the noted effects are due to so few independent variables.

For the institutions that were able to consistently maintain financial viability the dependent variable, full-time equivalent student, was positively related to the two independent variables, constant dollar net expenditures per student and acceptance ratio (adjusted $R^2 = .439$), as can be seen in Figure 7. This result implied that less selective entrance requirements led to a larger student population and so to a larger expenditure base. An alternative explanation for the acceptance ratio effect would be that the market niche of each of these institutions is clearly understood by potential students. The lack of a relationship with tuition discount percentage and constant dollar net student revenues might also suggest that, in this group, the larger institutions scholarship with restricted funds as opposed to leveraging with unrestricted current funds. For the institutions that were not able to consistently maintain financial viability the independent variables, tuition discount percentage and acceptance ratio, were positively related to the dependent variable full-time equivalent student ratio (adjusted $R^2 = .323$).

Figure 7. Significant results for the dependent variable: full-time equivalent student.

Strong group.

Regression Statistics	
Multiple R	0.687
R Square	0.473
Adjusted R Square	0.439
Standard Error	453.587
Observations	34

ANOVA					
	df	SS	MS	F	Significance F
Regression	2	5716405.062	2858202.53	13.892	0.000
Residual	31	6377963.398	205740.755		
Total	33	12094368.46			

	Coefficients	Standard Error	t Stat	P-value
Intercept	2884.694	495.518	5.822	0.000
Expenditures per Student	-0.348	0.067	-5.163	0.000
Acceptance Ratio	-869.179	777.033	-1.119	0.272

Weak group.

Regression Statistics	
Multiple R	0.599
R Square	0.359
Adjusted R Square	0.323
Standard Error	420.206
Observations	39

ANOVA

	df	SS	MS	F	Significance F
Regression	2	3556039.554	1778019.78	10.070	0.000
Residual	36	6356641.902	176573.386		
Total	38	9912681.456			

	Coefficients	Standard Error	t Stat	P-value
Intercept	-131.099	314.892	-0.416	0.680
Tuition Discount Percentage	1427.638	642.304	2.223	0.033
Acceptance Ratio	2458.980	1026.606	2.395	0.022

For this group, the implication is that the institutions with a larger student population accept more potential matriculants and leverage the cost to attend with unrestricted current fund dollars. Taken together, these two results clearly suggest that some combination of less selectivity or identification to market niche combined with a higher level of financial aid, or leveraged tuition, was related to a larger student population in both groups.

Discussion

Taken together, the results related to these performance indicators suggest that the recruitment and retention program is an important source of institutional financial viability. The results indicate that leveraging the cost to attend is integral to maintaining and expanding the student population for these institutions. The implication was that all the institutions discount tuition, though the financially more viable institutions were seen to rely less on discounting and more on funded scholarships. The performance indicators used for this study are among the most frequently cited as measures of institutional viability and the results did provide information related to institutional financial viability.

This study did demonstrate that there are unique groupings of liberal arts institutions and that unique financial equations for these groups might be defined in terms of several performance indicators. However, what does stand out is that there are few policy-related implications. These institutions, most of which have been in existence for over a century, are maintaining enrollment and graduating students. Some have more financial flexibility than others and that can be traced to size of enrollment and cost to attend. These standard financial ratios were being considered in a number of states as triggers for audits during deliberations related to the Statewide Postsecondary Review Entities (SPRE). Equations involving these financial viability indicators are being considered in the proposal to amend the Student Assistance General Provision regulations by revising the requirements for compliance audits, as detailed in the Federal Register Volume 61. Certainly these results did not imply that these frequently cited performance indicators should trigger federal policy and institutional sanctions.

The results of Study One did suggest that serious consideration should be given to questions of institutional viability, unique institutional profiles and the use of performance indicators in institutional management. Study Two explored a whole system approach developed around the concept of decision support as suggested by Kaufman in Educational System Planning (1972). One of the institutions included in the first study was used for the case study approach employed in Study Two.

Study Two

Review of the Problem and Literature

Institutions of higher education are, by any standard, complex entities. Even the least complex of institutions, the small liberal arts college, provides an enormous number of pedagogical, social, behavioral and economic phenomena to study. As campus decision-makers begin to understand these phenomena they become more effective at defining and creating the information needed to support decision making. The campus year might be envisioned as multiple threads woven together. Among these threads would be the recruitment and retention thread, an academic programs thread, a student life thread, a staffing thread, a physical plant thread, and a fiscal thread. Along each of the threads lie decision points. The sum of the decisions at these points are instrumental in creating the fabric and design of an institution's future.

It is a fairly straightforward task to list some of the critical decision points in the campus year and the questions they raise. What decision rule will we use for admitting students? How will financial aid be apportioned? Will there be unfunded financial aid, and if so how much? Will there be a raise? Can maintenance be deferred? What programs will be targeted for excellence and at what expense?

The answers to these questions, and a myriad more that confront the campus administrative and planning team will be cast in terms of decisions. At the very least, the leadership of every campus must ask the following two questions at the beginning of each academic and planning year. First, will we be intentional in making decisions for this campus? And, will we use the best possible information to reduce uncertainty before we make decisions? Assuming that decisions are to be intentional, our primary concern then is the need to reduce uncertainty before the decision is made. It is the role of institutional research to provide the information that reduces uncertainty prior to making decisions.

There are a number of decision points during the campus year encompassing a number of dimensions from departmental decisions to decisions with campus-wide implications. From a temporal perspective, there are decisions that are made daily, weekly, each academic term, and yearly. Almost all of the literature related to decision support focuses on the for-profit business and industry sector. This literature began to call for, and then examine, integrated decision support systems (DSS) starting in the early 1970s (Van Gundy, 1988). These analytical software engines were intended to provide the necessary decision support information at the appropriate desk for everything from daily to annual decisions throughout the firm (Alavim and Joachimsthaler, 1992). Implementation of completely integrated decision support systems in the for-profit sector has been marked by mixed results and the implementation of such systems remains a complex issue (Lucas, Ginzberg, and Schultz, 1990). The control of operations and support for marketing have seen a wide spread acceptance and use of decision support tools, primarily for daily, weekly and quarterly decisions (Alavim and Joachimsthaler, 1992). The literature on the use of decision support systems for major policy and direction related issues has shown that there is far less consensus on the use of DSS by top-level management (Reagan-Cirincione et al., 1991).

The acceptance of decision support systems in postsecondary education is similar to the experience of the for-profit sector, though the literature is not as rich. Most of the administrative software systems in use by the institutions provide adequate support for daily, weekly, and academic term decisions. The marketing function, embodied in the admissions and development programs, have become quite sophisticated. However, the use of information to support decisions related to the major policy, performance, and direction related issues faced by institutions leaves much to be desired (Gaither, Nedwek and Neal, 1994; Kidwell and Long, 1995).

For the purposes of this research, those decisions will be defined as decision points. Information developed to support those decisions is defined as a performance indicator (PI). For example, the decision to admit or not admit a student is, in fact, a daily or weekly decision. However, setting a decision rule that some measure, such a s school class rank, will be used as an admission criteria is probably done only once a year. This is a key decision point. The information used to make that decision, probably developed from a retention study and related descriptive statistics, would be defined as key performance indicators.

Though there is a large body of institutional research literature, that literature should be strengthened in three areas: 1. There is a need to develop a taxonomy of key decision points within the campus year; 2. There is a need to understand what key performance indicators reduce uncertainty prior to making a given decision and the impact of the information on decision making; 3. There is a need to understand the campus as a system defined by decision points and sets of decision points that are interrelated.

From a practical perspective, decisions are approached, and usually made, within the context of the institution's program structure. A framework for program structure was established nationally in the 1960s and has evolved into the current national program classification structure defined in NACUBO's

Administrative Service and implicit in the National Center for Educational Statistics Integrated Postsecondary Education System. Specific decisions will be made relative to the goals or budgetable objectives of a specific program or the cost centers defined at the sub-program or sub-sub-program level of the Program Classification Structure. Decisions are also within a temporal plane and related to specific times within the academic or fiscal year. A decision point is defined here as related to a specific program at a specific time in the academic or fiscal year. A decision model of the campus could be made that resembles a PERT chart with each line representing a program and the action points representing decision points.

Decision points can also be characterized in terms of the type of decisions that are made. Most will be regular and identifiable, located within the aegis of a program and at a specified time within the year or academic term. Other decisions will be unexpected and will encompass either new opportunities or decisions that need to be revisited. Decisions that need to be revisited are inevitable, even the best plans will require mid-course corrections.

Specific decisions are made and there are discreet decision points. However, decisions are rarely made in a vacuum. Specific decision points group together within decision sets. The information that is developed for the reduction of uncertainty at each decision point within a decision set is often reviewed together. Specific decisions are made within the context of the decision set.

If the most appropriate framework for decisions is the decision set, the nature of decision sets can best be described as a cascade. Even single decisions can lead to a cascade of additional discrete decisions. Multiple measures of outcome can be impacted in the same way. Perhaps the most important skill in policy analysis is being able to understand and predict the cascade effect.

This case study focuses on a campus that was included in the sample of institutions used above to explore the use of performance indicators as measures or predictors of institutional viability. In this particular case study the institution decided, in FY 1991- 92, that challenges on two fronts were threatening the institution. The first challenge was in the retention of students, with only 40% of entering freshmen returning for the second year. The institution was convinced that this was unacceptable in terms of cost to the institution to recruit a large freshman class, and in relation to the mission of the institution. The second challenge was in the construction of two new state supported branch community college campuses serving nearby counties. These counties had traditionally been a source of students for the institution, though many of these local students required remediation. The institution had a number of medial courses included within the academic program.

Decision Sets: A Case Study

Retention Challange

The first step was to collect information from the student record files and an entering freshman survey the institution had been administering and conduct a probability regression analysis to determine factors that correlate with retention into the second year. The results of that study are outlined in Figure 8.

Figure 8. Factors related to retention the second year.

Dependent Variable: First-time freshmen returning for second year.

Independent Variables Investigated with probit analysis

- ACT English
- Amount of Loans
- ACT Math
- Amount of Workstudy Hours
- ACT Social Studies or Reading
- Distance from Home
- ACT Science
- Dorm Student
- ACT Composite
- College Grade Point Average
- Graduation Quartile
- Gender
- High School Grade Point Average
- Married
- Amount Non-institutional Aid
- Elected Major
- Amount Institutional Aid
- Religious Preference

Groupings of variables found to significantly increase the probability of returning for the second year with probit analysis

Group One

- College Grade Point Average
- Dorm Student
- Institutional Aid
- None Institutional Aid
- ACT Composite

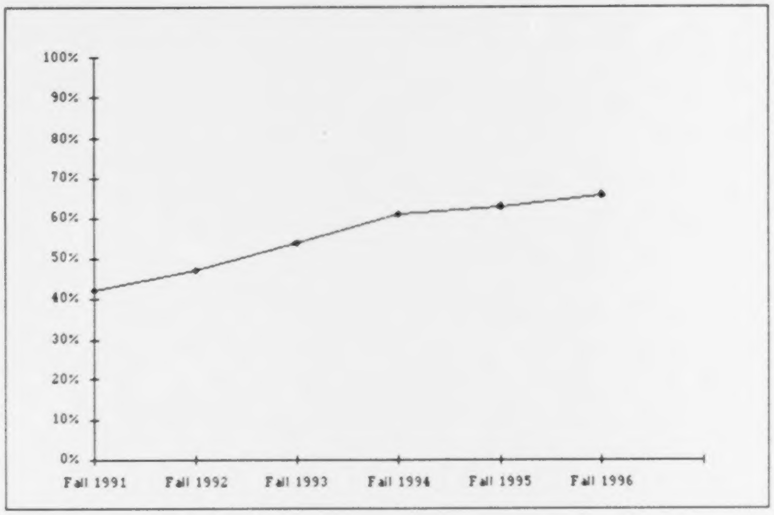
Group Two

- College Grade Point Average
- Elected Major
- High School Grade Point Average
- Dorm Student
- Loans

Given the results of the retention study, four policy related decisions were made.

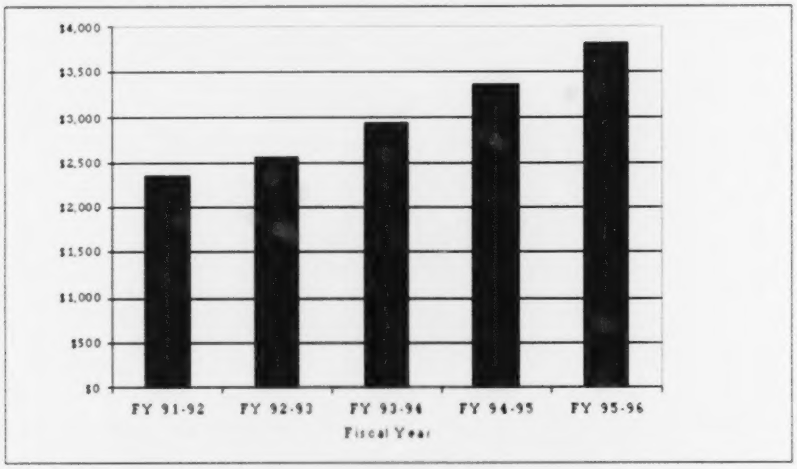
1. All freshmen were required to live in a college dorm except those living with a relative.
2. Admissions standards were refined and evaluation of all applicants was moved to a faculty committee using a multiple criteria best-fit model.
3. Policy for awarding financial aid was changed to focus on students most likely to be retained.
4. Faculty began to work with students on electing a major before arriving on campus.

Figure 9. Percent of freshman returning for the second year.



As can be seen in Figure 9 above the percent of freshmen returning for the second year rose dramatically from Fall 1992 to Fall 1996. Also, an intentional decision was made to increase the use of financial aid to recruit students who were more likely to be retained. Figure 10 shows the increase in scholarship and fellowship aid per full-time student from FY 1992-93 to FY 1996-97. As significant as these changes are, they should be explored within the context of a related, yet separate, decision set that was being addressed at the same time.

Figure 10. Scholarship and fellowship aid per full-time student.



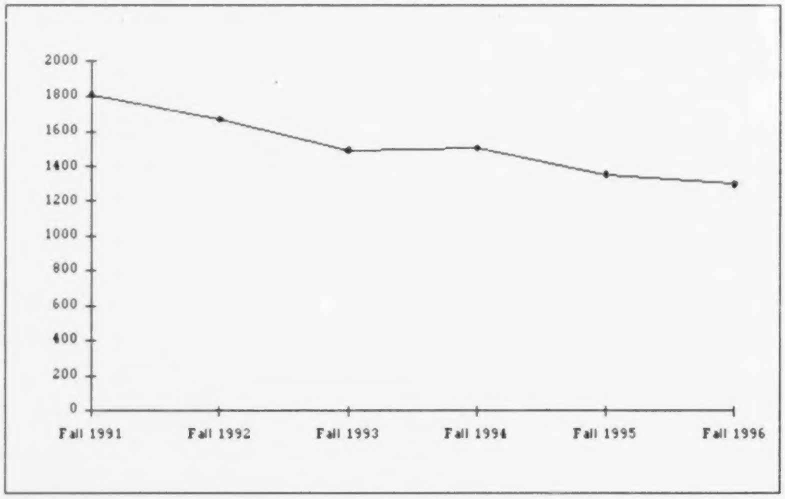
Community College Challenge

A significant segment of the institution's overall enrollment profile was from the counties surrounding the campus. Two community college branch campuses were opening in counties adjacent to the campus. Further examination of the student data set revealed that many of the students in the institution's freshman year remedial program were, in fact, students that would be candidates for these branch campuses and their open door policies. The institution made four key decisions:

1. to focus on students that have the highest chance of retention;
2. to withdraw from remedial programs and leave those students to community colleges;
3. to reduce the size of the student body concomitant with the withdrawal from remedial programs;
4. to reduce the size of the faculty relative to the size of the reduction in the student population.

As can be seen in Figure 11, the student population was significantly reduced between Fall 1991 and Fall 1996. There was an equivalent reduction in the faculty that occurred during that period of time. Faculty positions were reduced by 10% over the five year period with reductions related to remedial courses that were dropped. These faculty were in several disciplines and attrition and early retirement were the chief reduction in force strategies followed.

Figure 11. Fall enrollment 1991-1996.



A series of other decisions and results cascaded from these initial decisions. Some of these are outlined below.

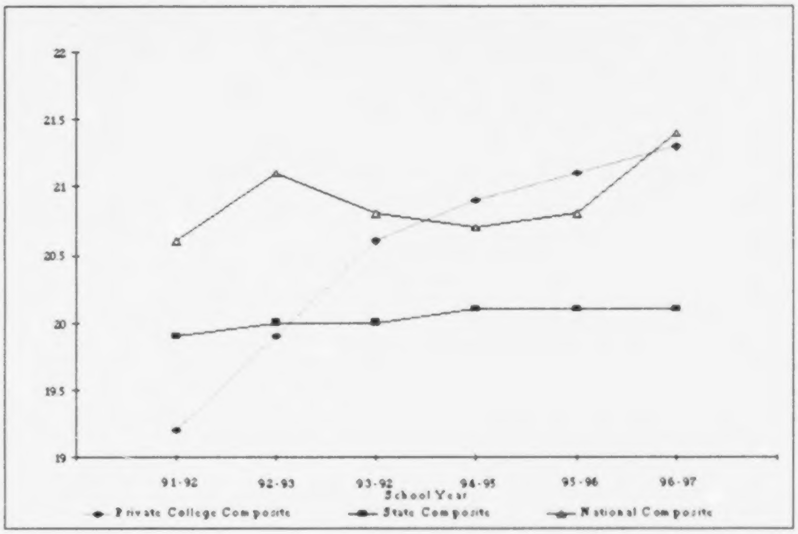
1. To attract and retain the caliber of student that had been identified as most likely to persist, the retention studies suggested some increase in financial

aid.

2. Though reduction in faculty offset most of the loss in student tuition and fee revenues, substantial increases in tuition and fees were necessary. Analysis had indicated that the institution was underselling its product.
3. The ACT scores of new freshmen increased dramatically as well as related measures of previous academic success.

Through intentional analysis and decision making the institution had changed the profile of its student body and reduced the size of the faculty. Though only one measure of the entering freshman class, the changing ACT profile, as seen in Figure 12, is indicative of the new more rigorous decision model for admitting students applied by the new admissions procedures.

Figure 12. Fall 1991-1996 new freshman ACT composite scores.



Perhaps the cascading nature of decision sets is also seen in the impact on the current funds. Figure 13 below shows the change in current fund expenditures as a percent of total from FY 1992-93 to FY 1996-97. The most significant feature of this period is the shift in expenditures from instruction to scholarships and fellowships. An additional impact is seen in reviewing tuition increases and the behavior of tuition and fee revenues at this institution during the five years being studied. Figure 14 indicates that tuition and fee revenue per student, net of scholarship, rose over the five-year period being studied. This was due to significant increases in tuition and fees and a decrease in number of students recruited. The institution was successful at recruiting a more academically prepared and affluent student population.

Figure 13. Current fund expenditures as a percent of total from FY 1992-93 to FY 1996-97.

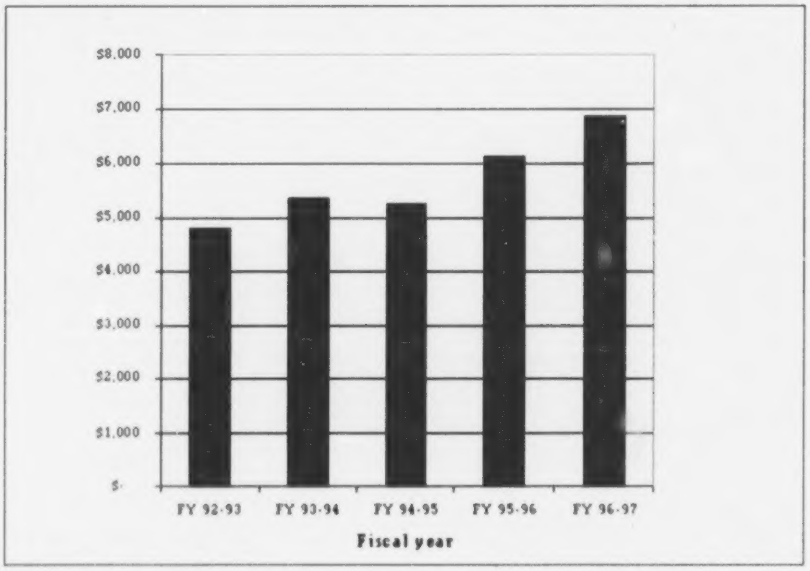
Expenditures by program

	FY 92-93	FY 93-94	FY 94-95	FY 95-96	FY 96-97
Instruction	$4,357,598	$4,452,393	$4,267,199	$4,343,119	$4,393,923
Academic Support	$734,794	$751,267	$757,292	$822,828	$887,878
Student Services	$3,209,702	$3,042,999	$3,686,927	$4,258,749	$3,648,876
Institutional Support	$2,742,729	$2,613,368	$2,581,843	$2,743,601	$2,477,247
Operation & Maintenance of Plant	$1,295,546	$1,445,542	$1,336,703	$1,556,446	$1,579,062
Scholarships & Fellowships	$3,512,193	$3,498,727	$3,278,937	$3,702,188	$5,512,980
Mandatory Transfers	$10,603	$11,971	$11,971	$19,031	$35,741
Auxiliary Services	$2,526,940	$1,909,342	$1,908,835	$2,297,643	$3,528,274
Total Expenditures	$18,390,105	$17,725,609	$17,829,707	$19,743,605	$22,063,981

Expenditures by program as a percent of total expenditures

Instruction	23.7%	25.1%	23.9%	22.0%	19.9%
Academic Support	4.0%	4.2%	4.2%	4.2%	4.0%
Student Services	17.5%	17.2%	20.7%	21.6%	16.5%
Institutional Support	14.9%	14.7%	14.5%	13.9%	11.2%
Operation & Maintenance of Plant	7.0%	8.2%	7.5%	7.9%	7.2%
Scholarships & Fellowships	19.1%	19.7%	18.4%	18.8%	25.0%
Mandatory Transfers	0.1%	0.1%	0.1%	0.1%	0.2%
Auxiliary Services	13.7%	10.8%	10.7%	11.6%	16.0%

Figure 14. Net tuition and fee revenue per student.



Discussion

This article approached the use of performance indicators from two perspectives. In the first study eleven frequently cited performance indicators were used to explore the implications of enrollment stability and financial viability with twenty similar Carnegie Classification Baccalaureate II institutions. This study examined issues addressed in the Federal Register Volume 61 proposal to amend the Student Assistance General Provision regulations by revising the requirements for compliance audits and adding a new subpart establishing financial responsibility standards. The implication here was that institutional scores on a specific set of indicators define the programmatic and financial viability of an institution and should impact the disbursement of federal funds. These results did not imply that these frequently cited performance indicators should trigger federal policy and institutional sanctions. What did stand out is that there are few policy related implications that can be drawn from these internationally accepted institutional viability measures. These institutions, most of which have been in existence for over a century, are maintaining enrollment and graduating students. Some have more financial flexibility than others and that can be traced to size of enrollment and cost to attend. What does stand out is that there are few policy-related implications.

The second study used a case study approach to focus on a campus included in the sample of institutions used in the first study. In this particular case study, the institution had decided that challenges on two fronts were threatening the institution. The institution moved to change both the population of students served and the focus of the academic program. The institution was successful over a five-year period in changing both the character of the student body and the academic program mix while improving its overall financial position. The institution used performance indicators within a whole system context, as suggested by Kaufman in *Educational System Planning* (1972), to reduce uncertainty before changing institutional policy and to measure the outcomes of those changes.

This case study was seen by the authors to reinforce their belief that specific decision points group together within decision sets. Information that was developed for the reduction of uncertainty at each decision point was reviewed, and decisions made, within the context of the decision set. The belief that decision sets exhibit cascade effects was also reinforced. In this case study single decisions led to a cascade of additional discrete decisions. As well, multiple measures of outcome were impacted in the same way.

Three overall conclusions were reached as a result of these two studies. First, the performance measures most commonly cited in the literature as measures of institutional financial viability are of limited use for institution specific policy development. Second, performance indicators are most effectively used within an institution specific, whole system framework. Third, being able to understand and predict the cascade effect in the use of performance indicators is essential for effective policy analysis.

References

Alavim, M. & Joachimsthaler, E.A.(1992). Revisiting DSS implementation research: a meta-analysis of the literature and suggestions for researchers. *MIS Quarterly, 16* (10) 95- 116.

Bogue, G. Creech, J. & Folger, J. (1993). *Assessing quality in higher education: policy actions in SREB States.* Atlanta, GA: Southern Regional Education Board

Dickmeyer, N.& Hughes, K. S. (1987). *Financial self- assessment workbook: a workbook for colleges and universities.* Washington, DC: National Association of College and University Business Officers.

Gaither, G., Nedwek, B.P. & Neal J.E. (1994). *Measuring up: the promises and pitfalls of performance indicators in higher education.* (ASHE-ERIC Higher Education Report; No. 5.) Washington, D.C.: The George Washington University, Graduate School of Education and Human Development.

Joint Commission on Accountability Reporting. *A need answered.* Washington , DC: American Association of State Colleges and Universities.

Lucas, H.C., Ginzberg, M.J. & Schultz, R.L. (1990). *Information systems implementation: testing a structural model.* Ablex Publishing Corporation, Norwood, NJ.

Kaufman, R. A. (1972). *Educational Systems Planning.* Englewood Cliffs, NJ: Prentice Hall

Kidwell, J.J. & Long, L. (1995). *Performance measurement systems for higher education.* (NACUBO's Effective Managment Series) Washington, DC: National Association of College and University Business Officers.

Notice of Proposed Rule Making, Number 184, Volume 61, Federal Register (1996)

Reagan-Cirincione, P, Shuman, S., Richardson, G.P. & Dorf, S.A. (1991). *Decision modeling: tools for strategic thinking*. Interfaces, 21:6, 52-65.

Ruppert, S. (ed) (1994). *Charting higher education accountability*. Boulder, CO: Education Commission of the States.

Prager, McCarthy & Sealy, Inc. (1995). *Ratio analysis in higher edeucation.* New York, NY: Prager, McCarthy & Sealy.

Sapp, M.M.(1994). *Setting up a key success index report: a how-to manual.* (AIR Professional File Number 51, Winter, 1994) Tallahassee, FL: Association for Institutional Research.

Taylor, B.E., Myerson, J.W. & Massy,W.F. (1993). *Strategic indicators in higher education: improving performance*. Princeton, NJ: Peterson's Guides.

Van Gundy, A.B., Jr.(1988). *Techniques for Structured Problem Solving*, second edition, Van Nostrand Reinhold, New York.

About the Authors

E. Raymond Hackett is Executive Director of the College Information Systems Association, a data sharing consortium of liberals arts colleges and universities. He is also an Assistant Professor in the Educational Leadership program at Auburn University. Dr. Hackett has served as a state system-level policy analyst and as a campus vice president for administration and finance.
Email: hackera@mail.auburn.edu

Sarah D. Carrigan is Coordinator of Institutional Studies at the University of North Carolina at Greensboro and formerly served as the Research Director for the College Information Systems Association. Dr. Carrigan has also served as a campus residence life director at a liberal arts college.

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb of the University of New Hampshire: casey.cobb@unh.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

John Covaleskie
Northern Michigan University

Alan Davis
University of Colorado, Denver

Mark E. Fetler
California Commission on Teacher Credentialing

Thomas F. Green
Syracuse University

Arlen Gullickson
Western Michigan University

Aimee Howley
Marshall University

William Hunter
University of Calgary

Daniel Kallós
Umeå University

Thomas Mauhs-Pugh
Rocky Mountain College

William McInerney
Purdue University

Les McLean
University of Toronto

Anne L. Pemberton
apembert@pen.k12.va.us

Richard C. Richardson
Arizona State University

Dennis Sayers
Ann Leavenworth Center
for Accelerated Learning

Michael Scriven
scriven@aol.com

Robert Stonehill
U.S. Department of Education

Greg Camilli
Rutgers University

Andrew Coulson
a_coulson@msn.com

Sherman Dorn
University of South Florida

Richard Garlikov
hmwkhelp@scott.net

Alison I. Griffith
York University

Ernest R. House
University of Colorado

Craig B. Howley
Appalachia Educational Laboratory

Richard M. Jaeger
University of North
Carolina--Greensboro

Benjamin Levin
University of Manitoba

Dewayne Matthews
Western Interstate Commission for Higher Education

Mary McKeown-Moak
MGT of America (Austin, TX)

Susan Bobbitt Nolen
University of Washington

Hugh G. Petrie
SUNY Buffalo

Anthony G. Rud Jr.
Purdue University

Jay D. Scribner
University of Texas at Austin

Robert E. Stake
University of Illinois--UC

Robert T. Stout
Arizona State University

---

archives | abstracts | editors | board | submit | comment | subscribe | search

EPAA • enter the archives • browse the abstracts • the editors • the edit board
• submit article • submit commentary • search • subscribe
volume: • 1 • 2 • 3 • 4 • 5 • 6 • 7

This article has been retrieved 290 times since September 9, 1998.

Education Policy Analysis Archives

Volume 6 Number 18 September 8, 1998 ISSN 1068-2341

A peer-reviewed scholarly electronic journal.
Editor: Gene V Glass Glass@ASU.EDU.
College of Education
Arizona State University,Tempe AZ
85287-2411

Copyright 1998, the EDUCATION POLICY ANALYSIS ARCHIVES.Permission is hereby granted to copy any article provided that EDUCATION POLICY ANALYSIS ARCHIVES is credited and copies are not sold.

The Transformation of Taiwan's Upper Secondary Education System: A Policy Analysis

Hueih-Lirng Laih
Su-Lin, Taipei County
Taiwan

Ian Westbury
University of Illinois at Urbana-Champaign

Abstract This paper explores the policy issues circling around the structural "transition" in upper secondary education implicit in the twenty-year increase in secondary and third-level school enrollment rates in Taiwan. This expansion has taken place within a secondary school system which is rigidly divided into both general, i.e., academic, and vocational tracks and into public and private sectors: the majority of students are enrolled in the private vocational sector which is only loosely articulated with the university sector. These features of the school system are analysed against the background of social and economic developments in Taiwan as well as public opinion. The analysis suggests that the present structures of school must be "reformed" in ways that will result in a more unified secondary system with both greater public funding and better articulation of all school types with the third level. The policy options that circle around the possibility of such reforms in the areas of curriculum, examination structures and second level-third level articulation are discussed and a policy framework for the reform of the Taiwan secondary education sector is outlined.

> My elder daughter is attending cram school to prepare for the two-year junior college entrance examination. (She didn't do well last year when she graduated [from high school].) It costs a lot of money to pay for the cram school, but I will do my best to support my daughter. It is encouraging to see her studying so hard. I wish my son was as ambitious as my daughter; he graduated from a public engineering [vocational] school three years' ago. He just didn't like school at all. But I think it would be better if he could stay longer at school to get more education. A decent job is not easy to get with a high school level diploma nowadays, is it?
>
> Mother of a middle school student

> I think there is too much difference in tuition fees between public and private schools. I believe private schools charge too much. This leaves the poor less choice in getting a proper education. From a taxpayer's point of view, the difference should be much less. That means the government should get involved by giving private schools more funds so that private school tuition can be reduced.
>
> Factory worker

The dramatic economic development and social modernization of Taiwan has, needless to say, been accompanied by increasing participation in the formal educational system, particularly at the secondary and third levels (see Figure 1). Between 1976 and 1995 net enrollment rates (including part-time students) in upper secondary schools (15-17 years of age) increased from 43 to 79 percent while net third-level enrollment rates (ages 18- 21) increased from 10 to 28 percent (Ministry of Education 1996: Table 4). (Note 1) Overall upper secondary level enrollment rates at these levels place Taiwan in the second tier among industrialized countries, along with the United States, Canada, and the United Kingdom, and ahead of Australia, Greece, and Spain (OECD, 1993: Table P13).

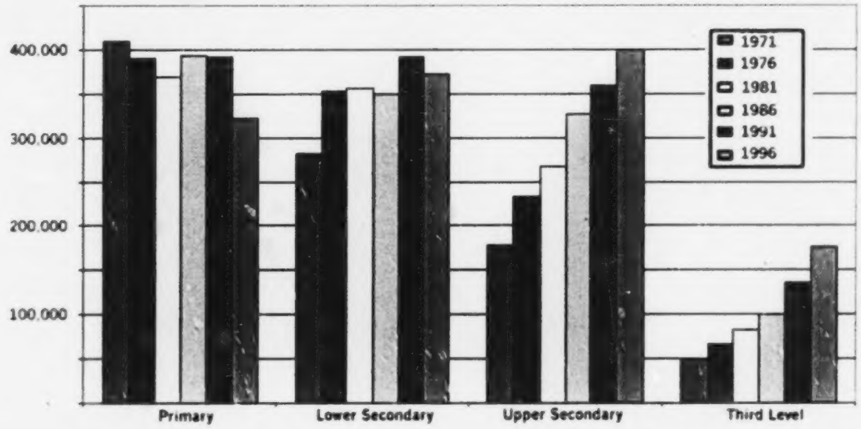


Figure 1: Numbers of students at all levels (1970-1995).

Our goal in this article is to explore and spell out the policy issues which we see circling around the structural "transition" in upper secondary education that is implied by the twenty-year increase in enrollment rates in this sector in Taiwan. As we attempt this, we will discuss, first, the background of these issues in the structures of Taiwan's present upper secondary and third-level systems. We will then consider some of the larger social and cultural forces which play on the expansion of Taiwan's secondary-level and college-level systems as a context for an examination of the pressures within and around these systems. Finally, we will speculate on the policy problems emerging from these pressures that need to be faced by Taiwan's educational policy makers.

As a background to this discussion--and we will be exploring these issues in detail below--we should note that the institutions that provide secondary and third-level education in Taiwan are divided into firmly separated academic and vocational tracks. It is a dual-track system. Thus, structurally, Taiwanese secondary education conforms to the pattern commonly found in continental Europe but unusual in English-speaking countries. In addition, Taiwan provides much of its secondary schooling, and particularly its vocational schooling, by way of a private sector which, while heavily regulated, receives only limited state support. Third-level education has a parallel structure with, again, a significant private sector.

In this article we will argue that these characteristics of the secondary and third-level systems pose, and will pose, major problems for educational policy makers as Taiwan's educational development continues. We will argue that the present pure dual-track, public/private system is, and increasingly will be, unable to accommodate the expectations for educational opportunities of Taiwan's families and youth--and will, therefore, require major "reform." This is increasingly recognized by Taiwan's policy makers and élites; however, we will also be arguing that the structures of the present system, and particularly its heavy reliance on the private sector, will make "reform" of the system quite difficult--and this is not so widely recognized.

The formal structures of the Taiwanese education system

Academic and vocational education

Figure 2 presents a schematic outline of the formal structures of the Taiwanese education system. This structure emerged after reforms in 1968 when the then six-year span of compulsory education was extended to nine years and, under the manpower- development economic policies of the then-government, the increasing number of students making the transition to upper secondary school were directed to the secondary vocational rather than the academic sector. (Note 2)

These general policies and the institutions for schooling that emerged from them have remained in place since the early 1970s to provide the framework for Taiwan's present secondary education system--with the consequences seen in Table 1. The percentage of students enrolled in the academic high school (full-time and supplementary) has dropped from 35 percent of the in-school cohort in 1971 to 20 percent in 1996 while vocational secondary (full-time and supplementary) and junior college enrollments have increased from 65 percent of the cohort in 1971 to 80 percent in 1996.

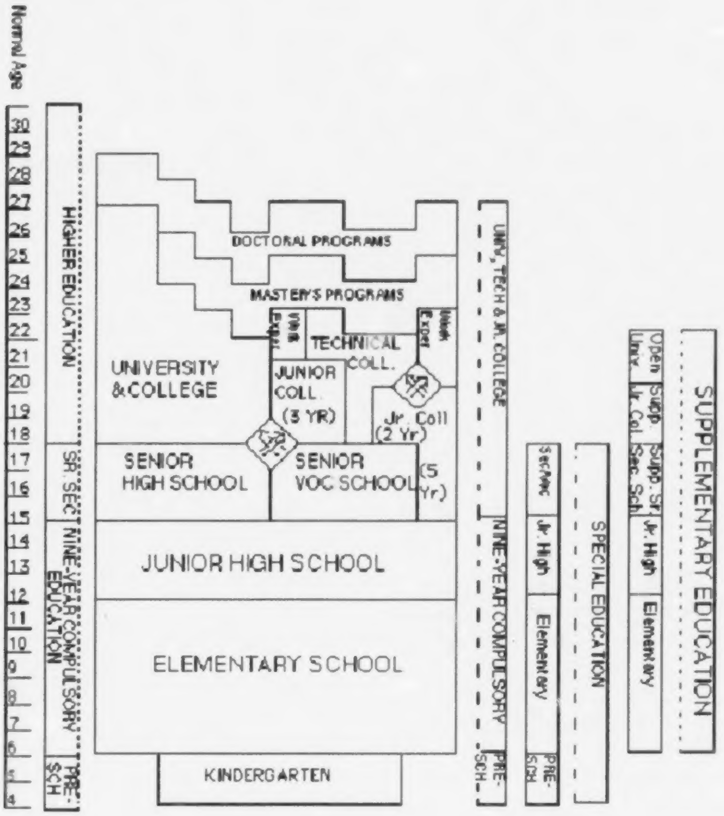

Normal Age
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
HIGHER EDUCATION
SR. SEC
NINE-YEAR COMPULSORY EDUCATION
PRE-SCH
DOCTORAL PROGRAMS
MASTER'S PROGRAMS
UNIVERSITY & COLLEGE
TECHNICAL COLL.
JUNIOR COLL. (3 YR)
Jr. Coll (2 Yr)
SENIOR HIGH SCHOOL
SENIOR VOC SCHOOL
(5 Yr)
JUNIOR HIGH SCHOOL
ELEMENTARY SCHOOL
KINDERGARTEN
UNIV, TECH & JR. COLLEGE
NINE-YEAR COMPULSORY
PRE-SCH
Sec/Voc
Jr. High
Elementary
PRE-SCH
SPECIAL EDUCATION
Open Univ.
Supp. Jr. Col.
Supp. Sr. Sec. Sch
Jr. High
Elementary
SUPPLEMENTARY EDUCATION

Figure 2: Structures of schooling in Taiwan.

The implications of these enrollment patterns in upper secondary schools must also be considered in the light of both the growth in third-level enrollment and the structures of articulation of the different secondary schools types with the third-level system. Thus, in recent years the third level has seen the same pattern of growth as the secondary level: as we noted above, net enrollment rates in third-level institutions have increased from 11 percent of the population aged 18-21 in 1980 to 28 percent in 1996. However this overall enrollment rate conceals substantial differences in the transition to the third level by graduates of secondary academic and vocational schools (Ministry of Education, 1996). About 60 percent of the students graduating from academic high schools entered the third level in 1992 as compared to 20 percent of vocational graduates (Department of Education of Taiwan Province, 1994c). (Note 3)

Table 1
Student Enrollment (Percent) in Upper Secondary Level by School Types (1971-1996)

Year	Academic High School	Vocational School	Junior College	Supplementary Vocational School	Supplementary Academic High School
1971	33.1%	43.0%	11.7%	10.4%	1.8%
1976	23.0	50.0	11.0	14.6	1.3
1981	20.9	52.9	12.5	12.6	1.2
1986	17.9	52.9	12.4	15.8	0.9
1991	17.9	51.7	15.7	14.1	0.6
1996	20.1	49.7	18.1	11.7	0.4

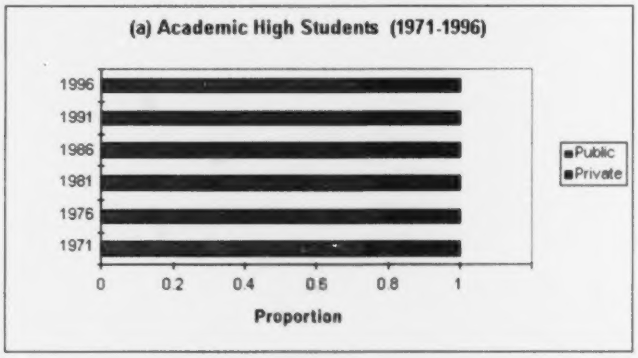
Source: Ministry of Education. 1997.

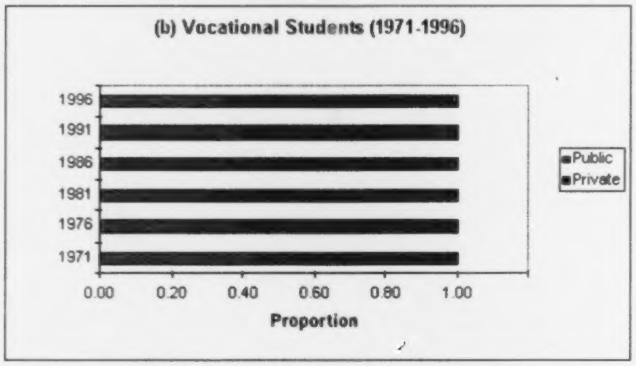
In other words, the possibility of successfully transferring to the third level--with all of its status and occupational opportunities--is tightly linked to secondary school track. This arises from both the organization of institutions within the third level and the way in which access by students to the third level is organized. Entry from secondary school to third-level institutions is mediated by a set of examinations: academic high school graduates take one of set of examinations based on the prescribed curriculum of the academic secondary school; success on the examination determines which program and institution a student will be admitted to; vocational school graduates take a different set of examinations which are again based on the (different) vocational school curricula which allocate students, depending on achievement, to the institutions (i.e., two-and three-year junior colleges) which are formally articulated with the vocational sector. Only limited transfer from the secondary vocational sector to the academic third-level sector is possible at the point of entry to the third level.

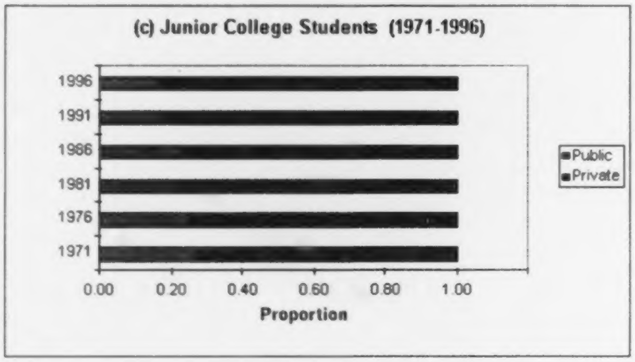
Public and private schooling

In addition to the structural differences between school types or tracks, Taiwan has, as we noted earlier, a mixed, private- public pattern of educational provision--with a substantial private sector, particularly in the secondary and third-level vocational sectors.

Figure 3 presents the proportion of private and public places in the upper secondary school types and university sector since 1970. The state has provided the bulk of the senior high school places (more than 70%) for most of the period; however in the secondary vocational sector (both vocational high school and junior college) the role of the state shrinks dramatically and has declined over the period, from state provision of about 53% of the places in 1970 to about 37% in 1996.


(a) Academic High Students (1971-1996)


(b) Vocational Students (1971-1996)


(c) Junior College Students (1971-1996)

BEST COPY AVAILABLE

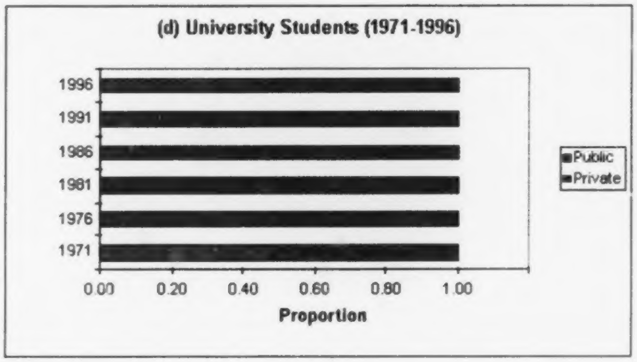


Figure 3: Public and private enrollments in upper secondary school types and universities (1970-1996).

Private secondary-level schools receive only limited support from the state. The result is that private schools assess much higher school fees than do public schools but the per capita expenditures in these schools are much lower than in public schools--and per capita expenditures have increased more in public than in private schools. (Note 4) These differences reflect the cost structures of a private schools, but, as Chen (1993) reports, reflecting Ministry of Education findings, they are accompanied by a lack of investment in facilities and sub- standard equipment.

In summary, much of the responsibility for the provision of places to accommodate the expansion of demand for secondary schooling, i.e., in vocational schools, has been given to the private sector. At the same time, the state has provided significant subsidies to the smaller number of academic and vocational secondary students enrolled in state-sponsored schools. (Note 5) We cannot here consider the historical roots or concomitants of these priorities (Note 6) but, as we argue later, their present consequences pose real problems for the transformation that we believe that Taiwan's secondary education system must undergo over the next 20 years. But before considering these issues let us consider some explanations of the forces underlying the increasing commitment to schooling as the pathway to adulthood that is occurring in Taiwan. The dynamics that these explanations provide a firm basis for foreseeing the problems and tensions that the system will face over the next decade.

The expansion of the role of school in the pathway to adulthood

It is a commonplace that schooling has assumed a dominating role in the pathway to adulthood in modern societies. To understand this expansion of the school's role we must consider, first, its consequences and, second, its causes. The consequences associated with this expansion involve fewer explanatory issues than do its causes.

Thus, Trow (1960) captured many of the implications of school (and college) expansion, both for the changing social roles and the educational characters of the secondary schools, with his now-classical characterization of the stages of the American secondary school's movement towards the hegemony of the school as an institution dominating young adulthood. At the beginning of the process, in the pre-world war 1 period, the American secondary school was an *élite-preparatory* institution enrolling a small proportion of the age cohort and offering a curriculum that assumed that many of its graduates would, or should, advance to some form of higher education. This school was succeeded in the inter-war period by a *mass-terminal* school with a significant vocational orientation and curriculum in which there was widespread participation to the end of a secondary education, but most students did not continue their school careers after this point. This mass-terminal school changed in the post-World War II era to the *mass- preparatory* school in which the college-preparatory curriculum again assumed major importance for the secondary school, although the terminal function continued for many. While the specific terms of Trow's account of the transition of the secondary school were embedded within the particular transformation of the U.S. high school in the 1950s, his framework has been seen to have a world-wide validity, even if its expression might differ across societies.

But how can we explain this increasing hegemony of the school as an institution over the pathway to adulthood? It has been usual to attribute the dominance of, first, the mass-terminal secondary school and, later, the mass-preparatory secondary school with its link to the modern college and university to the need for the specific forms of human capital required in modern economies that, it is assumed, schooling alone can provide. However, it is, as Dreeben (1972) argues, not self-evident that the hypothesis of such a linkage can be sustained. As he argues, perfectly adequate occupational training of every kind has been, and is, provided through apprenticeship broadly conceived. It is, for example, not clearly the case that those who prepared for legal careers via articles, i.e., apprenticeship, are or must be less skillful than law school- trained attorneys. Or that graduate education of physicians makes for unambiguously "better" physicians than undergraduate medical training. Furthermore, while there have been and are careers that are intimately associated with advanced schooling, many of these careers have been and are within the (expanding) institution of schooling itself, i.e., teaching in schools or universities. Indeed, as Dreeben (1972) points out, it is only teaching as an occupation that can be seen as clearly associated with the expansion of schooling!

In the face of such difficulties with using human capital arguments to account for the expansion of the scope of schooling, Dreeben makes a different case for the "success" of the school as an institution in modern societies. He links schooling, first, to its most basic role in communicating broad literacies and, second and following Marshall (1964) and Parsons (see Englund 1996), to its symbolic and institutional role as a concomitant of an expanding conception of "citizenship," with its accompanying rights. In the course of this century in the United States and, more recently, in most other industrialized societies, advanced (secondary and third-level) schooling, like health care, state-provided welfare, and other income-transfer programs more generally, has become integrated into social understandings of the rights of access to valued social goods associated with the idea of citizenship. In this analysis Trow's account of the expansion of the scope of the American secondary school--and the related expansion of the scope of the college-- becomes a manifestation of an expansion of the nature and scope of the idea of a common and universal citizenship seen as the right to participate fully in the institutions of the social and cultural order.

This right of and demand for education is, of course, exercised through a particular interactions between the ambitions and capacities of families, students, etc., and corporate actors, such as the state, which both provide schooling either directly or by way of subsidy or legitimation of private providers and also define the framework of occupational credentials and the forms of rationing of these credentials, *not* of occupational skills as such. The ways in which such interactions play themselves out vary, of course, depending on the characteristics of particular regimes. But, in general, public demand for education and/or credentials induces the state both to create institutions to meet those needs and to regulate their availability--because the very legitimacy of the state requires responsiveness to both "public" and "special" interests (Craig, 1981; Craig & Spear, 1982a, b). (Note 7) This, in its turn, channels societal expectations, and thus "public" interests and the interests of a regime, as a provider of education and the legitimator of the credentials, etc., converge. Schooling becomes *the* pathway to adulthood because of its legitimation of occupational credentials as well as the rationing of the availability of these credentials, *not* of occupational skills as such. And, of course, this convergence is most complete when a state is, or claims to be, fully democratic, i.e., responsive to its citizens and their interest groups.

It is forces such as these which are currently working themselves out in Taiwan as both socio-economic and political developments converge.

Social development and citizenship rights in Taiwan

There is no need to repeat here the story of economic and social development in Taiwan over the past three decades. High growth rates of GNP have resulted in incomes and living standards that have reached levels which are comparable to those of western industrialized nations. This development has, moreover, been experienced by much of the population--with the result that Taiwan has one of the most equitable distributions of wealth among both developing and industrialized countries (Deininger & Squire, 1996). As a result, a substantial "middle class," defined fairly narrowly, has emerged with an estimated size of between 25 to 40 percent of the adult population (Tien, 1989, p. 33; Tien, 1992, p. 36).

These economic changes have inevitably led to changing socio-cultural *perceptions*. As seen in Table 2, as long ago as 1982 36 percent of Taiwanese saw their parents as middle class but 54 percent saw themselves as middle class. Nine percent saw themselves as upper middle class--but 33 percent saw their offspring as becoming upper middle class (Cheng, 1993).

Table 2
Perceptions of Social Stratification in Taiwan

Social Stratum	Parents	Self	Offspring
Upper	1.2%	0.6%	7.3%
Upper-middle	7.2	9.3	32.6
Middle	36.1	54.7	36.9
Lower-middle	36.4	26.8	5.8
Lower	19.1	8.6	1.0
Uncertain	-	-	16.4
Total	100.0%	100.0%	100.0%

Source: Cheng, 1993.

Such social and cultural changes have pushed and made possible Taiwan's political and educational transformation. The long- term ruling political party, the Kuomintang (KMT), now retains political power on the foundation of a real election victory and is, moreover, "investing heavily in . . . policy areas where the general public has an immediate . . . stake. These areas include social welfare, environment, consumer welfare, regional development, and many other issues that are common to a society reaching a higher stage of economic development" (Cheng, 1993: 214). The consequences of these developments, when linked to the traditional Chinese commitment to and respect for formal academic education, have (and will have) profound implications for both social demand for secondary and third-level education and for understanding about how it should be provided--in terms both of the state's role as a provider and institutional frameworks of provision and credentials.

We can detect traces of this social and cultural demand in several educational indicators. Thus about 50 percent of middle school 1st graders (grade 7) hope to enter an academic high school and 35-40 percent of 3rd graders (grade 9) have the same aspiration. Sixty to 70 percent of middle school students in the capital, Taipei, as distinct from 40-45 percent of middle school students in Taiwan Province, plan to enter an academic high school (Ministry of Education, 1994). The transition rate of junior high school graduates entering senior secondary school increased from 68 percent in 1981 to more than 80 percent in 1988 and about 90 percent in 1996 . Senior secondary school net enrollment rates increased from 53 percent in 1981 to 73 percent in 1990 and about 80 percent in 1996. Between 1988 and 1992 the transition rate of senior secondary graduates entering third-level institutions increased from 19 to 31 percent.

However as we have already noted, such aggregated data, with its clear evidence of an increasing commitment to schooling as a pathway to adulthood, conceals major differences between the *opportunities* associated with the different secondary school types. Thus while the transfer rate of senior high school graduates to the third level increased from 45 to 57 percent between 1980 and 1995 that of vocational school graduates increased from three (in 1987) to only 20 percent. Only about five percent of vocational school graduates are admitted to universities (Department of Education of Taiwan Provincial Government 1994a, b), with the result that 90 percent of university students are graduates of academic high schools.

As we have also noted, Taiwan deploys secondary education by way of a complex, relatively "pure" multi-track system: enrollment in the two major school types at the secondary level, the general or academic high school (enrolling about 20 percent of students in 1996) and the vocational school (80 percent), represents very different educational opportunity structures and, in so doing, foreshadows very different--and increasingly different--educational and occupational careers. One part of the system has become, to use Trow's (1960) terms, a mass preparatory system while the other part remains a mass terminal system. In this system vocational students are severely disadvantaged--both in terms of their access to the full range of third-level opportunities, i.e. to the university sector, and the private costs (when compared to public, largely academic schools) associated with enrollment in a (in many ways) less desirable sector. It is this structural problem, and more important the institutions which flow from these structures, e.g., the examination systems which allocate students to upper secondary school types and the university sector, which creates many, although as we will see later not all, of the pressures and tensions the system is experiencing.

Defining the mismatch of supply and demand

What is the extent of the mismatch of the mismatch between the supply of places in the third-level university sector and demand for those places among vocational students? We can go some of the way in specifying its present scale using proxy data. Thus one estimate of the numbers of "dissatisfied" students in the vocational sector can be secured from registrations for the Joint College Entrance Examination. In 1994 approximately 12,000 of the 125,000 students registering for this examination were from vocational schools--although they had little chance of success. (Note 8) In 1993 136,808 students also registered for the College Transfer Examination; 9,006 of these registrants were admitted to colleges, 8,202 of whom were junior college graduates (Council of Educational Reform, 1995).

Another perspective on the size of the vocational school population aspiring to enter a university can be seen in the number of vocational school graduates who are not attending third-level institutions and not working; these missing persons are assumed to be attending a cram school to prepare for a university or junior college entrance examination (Department of Education of Taiwan Province 1994b). This group increased from 10 to 20 percent between 1977 and

1992, and it is estimated that about 80 percent of the group are planning to take a third-level entrance examinations after a year in a cram school. Extrapolating from the size of the graduating cohort, this suggests that there are currently about 23,000 vocational students actively aspiring to third-level entry.

Aggregate data on this kind gives one kind of picture of the "demand" for third-level places by vocational students, at least insofar as an estimate can be derived from actively "dissatisfied" upper level vocational students. But what of the silent majority of vocational students and the parents of those students? What are their attitudes toward their secondary school options? We sought to secure an understanding of these issues by way of face-to-face discussions with middle and vocational school students and their parents.

Four sets of middle and vocational school students in Taipei city and suburban Taipei and in a small community (population 45,000) in the southern part of the island were group- interviewed. We followed up these discussions with individual interviews of those parents of these students who we could contact either by phone or face-to-face. Both students and their parents came from both working and middle class families. (Note 9) Altogether 47 students (17 from vocational schools; 30 from middle schools) and 14 parents were involved in these discussions. While our sample of students and parents was opportunistic, it was not (we believe) biased in terms of social class or ethnicity.

Our goal in undertaking these discussions was to tap the feelings and attitudes of "typical" students and parents towards the systemic problems our more formal analysis seemed to be identifying. Thus we were concerned particularly with the views of, first, vocational school students and their parents towards the vocational sector of the secondary system and the inequitable opportunities for access to the third level that we saw it offering. We wondered how students and parents saw these issues. Second, we were interested in the views of middle school students and parents who were facing the issue of choice of a school type on these same issues. Overall our questions were: Do the typical clients of the system, and in particular the clients of the vocational sector, share in the understandings of the system--and the implicit critique--that emerges from an analysis of the kind that we were undertaking? Can we see evidence of an increasing dominance of the idea of schooling as *the* pathway to adulthood and a press towards the "academic" sector? Our analysis predicted such a movement in public attitudes towards schooling itself along with an increasing rejection of the vocational school based on our readings of theories of schooling expansion such as Trow (1960), Dreeben (1972), and Craig (1981; Craig & Spear, 1982a, b) which are, of course, largely based on American and western cases.

As will be seen below, the findings of the group interviews modulate and qualify--but also extend in one important way-- the interpretation of the major problems facing Taiwan's secondary and third-level educational system that we have been offering in this paper. They highlight the commitment to schooling that we would expect to find as well as a widespread understanding of the issues around the examination system that have been the focus of the most intensive policy making in recent years (Laih, 1995). However they also make clear that, although the academic high school was firmly perceived by those we interviewed as the preferred school type, most of the students and parents we interviewed did have positive attitudes towards vocational education as an option for themselves or their children. It was seen as offering a schooling that could provide useful practical preparation for work, although parents judged the secondary school as more desirable overall for their children if they could have the best of all worlds. Unexpectedly, it was the relative costs of public versus private secondary education that clearly emerged as a major concern for both students and parents. We had not anticipated the force of this attitude and feeling, although in hindsight such a view is consistent with the thrust of our understanding of the welfare-orientation that is accompanying the social and political changes taking place in Taiwan.

We summarize the themes that emerged from the interviews under three heads:

- attitudes towards the place of the third level in educational careers;
- the issues that are seen as circling around the choice of a secondary school that must be made after middle school; and
- attitudes towards the system of public and private schools--and their relative private costs.

We will let our informants speak in their own words.

Attitudes towards the third level

> I would like my two daughters to receive more education after they graduate from school [both are vocational school students], but they don't seem interested. They told me that they might do it after a few years' working experience; they want to experience life and see the world outside the school first. I can't say it is a bad idea, but what worries me is that they will finally find out how important it is to get a higher level of education. And if they do, it will be very difficult for them to pass the examinations then, because after years after leaving school, they will need to pick up all the subjects they learned in school. And, as you know, it is very difficult for fresh graduates to pass the examination.
>
> Mother of two vocational school students

> If my children were able to, going to a university is of course better--as everybody knows. Since they cannot, receiving education with job training is also a fine idea, but I think two more years' education after the vocational school is important in finding a better job.
>
> Mother of a middle school student

Academic versus vocational schools

We noted above that both the students and parents we interviewed had positive attitudes towards vocational education and were, in the main, satisfied with their schools they attended or planned to attend. At the same time, however, the differences between the school types and the connection between school type and the important issue of access to the third level were seen quite clearly by both students and parents. In the words of two of middle school students:

> My parents said it would be better if I can get into a academic high school, but they also said that, unless I can get into a college from there, it is useless. So I think it is a good idea to go for the vocational school; at least, I can avoid taking another entrance examination--which is like hell to me.
>
> Middle school girl

> Going to the academic high school is of course better because it is the way to go to universities; but it might not be as useful as the vocational school if you are not able to pass the college entrance exam. I know the exam is very difficult to pass, so I think settling for the vocational school is just fine for me.
>
> Middle school girl

> I think vocational school is more fun than the academic high school. I don't regret coming here. However, it is a fact that vocational students' chances of getting into advanced level are much less than the academic high school students. While our chance is below 20 percent, their chance is about 50 percent. I think it is not fair.
>
> Girl in a public vocational school

The sense of inequity stated in this last quote was not directly articulated by many students and parents in our sample. However, for some parents and students, the differences between public and private schools did raise another aspect of the issue of equity--indicating their understanding of the emerging issue of "equal" citizenship.

Public versus private school

As we noted above, the theme of the relative costs of public versus private schools was consistently introduced into our discussions with both parents and students as *the* immediate issue around secondary education. It was seen as a problem across social groups; it was also an issue which had a clear focus in that it was seen as an important target of potential government action.

> Some of my neighbors' sons and daughters will have to attend the supplementary school in the evening so that they can work during the day time to earn money to pay for the tuition. Although I and my husband can afford my daughter's tuition, I think the amount is really too high; it is really a problem for poor families.
>
> Wife of a proprietor of a small factory

> I think the government should give greater subsidies to the private school so that we can pay less tuition.
>
> Worker in small factory

Discussion

The troubling issues circling upper secondary and third-level education in Taiwan have been most often seen as centering on the mechanisms of allocation between school types, i.e., the examination system, and the attendant stresses this system places on students. These issues have been the focus of the most active recent policy initiation and policy making around upper secondary education. However when, in 1995, Yuang-Ze Lee, the highly regarded president of the Academica Sinica, initiated a campaign to abolish the current examination system, he introduced a new theme into policy discussions by noting that a necessary part of such reform would be the abolition of the distinct vocational schools by their transformation into general or academic high schools (Freedom Daily Tribune, March 23, 1995). The major opposition party, the Democratic Progressive Party (DPP), echoed and extended this argument by proposing the introduction of a comprehensive high school in its platform for the 1995 mayoral election in Taipei. The entry of such arguments about the structure of the school system into political debates in what is now a responsive polity foreshadows major changes in educational structure--but, as we will argue, major institutional changes in Taiwan's educational system will not be easily implemented. The long-standing structures of the secondary and third-level systems--with all of their cultural meaning--will pose major obstacles for such reforms.

Thus while a case can be made for the necessary change in the examination system, a focus on examination reform alone misreads the core problem facing Taiwanese secondary education. High school and university "entrance" examinations are only mechanisms for the controlled allocation of students to individual schools and school types, i.e., they are mechanisms for rationing access to scarce places. The mechanisms might be changed in any one of a number of ways, but the "problem" facing Taiwan's policy-makers would still remain. The examination issue merely serves to highlight the distribution of places in, and the structural rigidities of, the present multi-track post-compulsory education system with its secondary academic sector-- with its link to the university system--and the less desirable vocational sector--with its much weaker articulation with the third-level. It is the stratification of secondary education, and the increasing demand for the restricted but high-status "academic" track as this interacts with rising educational expectations, which will determine the future shape of Taiwan's upper secondary educational system. The issues which circle around the distribution of places among schools and school types are emerging as a major issue challenging educational policy making in Taiwan because of a potential fundamental transition of the secondary school from a mass-terminal to a mass-college preparatory institution. *Because of this secular change, we believe that the present structures, and the balance between the parts of the system, cannot remain in place.* But how can the system change? and what kind of change is forseeable?

As the analysis we have offered suggests, there are several, analytically distinct clusters of problems confronting the Taiwanese secondary and third-level systems. At one level there is a need for more places in the general or academic high school track to satisfy the increasingly widespread aspiration for both college-preparatory education. This issue intersects, however, with the larger problems circling around the framework for post-compulsory education with its interactions with the credentials and labor markets, which are themselves changing, so that it is not clear that any rigid (or "pure") binary system can provide either curricula that can embrace the larger numbers of students who aspire to higher-status post-compulsory education, or to an occupational preparation that meets the needs of changing, more knowledge-intensive credential and labor markets. There is, furthermore, the issue of the widely- perceived inequities associated with the public/private structure of provision with its state-funded high-status academic sector and the (largely) privately-funded lower-status vocational sector. The proposals of the reformers have called for a comprehensively organized state-based secondary sector; we must ask what this might mean and how it might be structured. What would be involved in any major policy shift away from the present pure, multi-track system, private/public to a different kind of system?

One way of framing these issues is suggested by Raffe's (1993) discussion of the issues circling around the reform of the Scottish post-compulsory and upper secondary system. In order to build his argument Raffe offers an analytical model based on a set of three ideal types of post-compulsory organization: a "pure" and a "flexible" "multi-track" organization and a "unified" organization although, as he notes, there is "no pure example of a unified system yet in existence" (p. 234). Table 3, taken from Raffe's paper, sets out the characteristics of each of these types and Table 4, also taken from Raffe's paper, describes the pathways that students may take in idealized pure and flexible multi-track systems.

Table 3
Types of Post-compulsory Education System

	Multi-track		**Unified**
	Pure	***Flexible***	
Basis of differentiation	Group	Group	Individual
Curriculum structure	Course, line, etc.	Course, line, etc.	Modular
Pathways	Limited transfer between tracks	Frequent opportunities for transfer	Flexible
Relation of stage to level	Relatively fixed within each track	Flexible	Flexible
Content	Academic and vocational	Academic and vocational, with large common element	Integrated/diverse
Certification	Separate systems	Separate systems with credit transfer	Single system
Principles of curriculum, pedagogy and assessment	Differ between tracks	Smaller differences	(More or less) consistent across system
Mode and institution	Differ between tracks	Flexible relation with tracks	Diverse, with dominant mode/ institution
Values	Differ between tracks	Possibly differ between tracks	Pluralist

Source: Raffe. 1993.

Table 4
Arguments for the Two Types of System

	Multi-Track System	**Unified System**
Curriculum and learning	• Academic standards • Vocational standards • Ability grouping • Avoiding ill-effects of modularization	• Integration of academic and vocational • Tailoring to individual needs
Incentives and Motivation	• Vocational/practical emphasis • Occupational identity • Avoiding academic drift	• Avoiding early "rejection" • Incentive of incremental decision-making • Incentive of mainstream certification • Avoiding credentialist pressure on academic track
Social	• Alternative criteria of success and esteem (horizontal differentiation)	• Later and less formal differentiation
Resources	• Avoiding costs of modular options	• Avoiding costs of separate tracks and specialisms
Coordination	• Competition among tracks	• Planning as coherent system

Source: Raffe. 1993.

Using Raffe's terminology, the present Taiwanese system clearly represents a pure, multi-track case with rigidly framed differentiation of clienteles, limited transfer between educational pathways, fixed staging, differentiated academic and vocational content, separate certification structures, different curricula, pedagogies, and assessment, different organizational delivery structures, and differentiated values. The outcome is a system which is experiencing significant stress because of the differential esteem and opportunities associated with the two sectors. These differences are, in their turn, associated with, and create, significant and widely felt problems for the middle school--at the end of which the decisive track determinations are made--and have stimulated the emergence of Taiwan's pervasive cram schools, which coach students for the high school and college entrance examinations.

As we have suggested, there is a strong basis for predicting that changes in the present system of educational service delivery in Taiwan are inevitable, and will be directed towards articulating the vocational school with the third level so that the vocational school can assume a clearer college-preparatory role. But what might the end-product of such changes look like? The present policy trajectory aims at expanding the articulation of the vocational system with an emerging multi-track third level by way of the degree-granting technical institutes (see Figure 2). In addition, moves are being made to expand the number of places in the academic secondary sector while maintaining the multi-track structure of the larger system. *But such reforms are essentially piece-meal in that they do not address the fundamental inflexibility of the present system's overall structures and the pervasive stresses around these structures, which derive both from the rigidities of the present multi-track system and the cost-differentials and expenditure- differentials between the private and public sectors.* Thus, at present the *only* direct and clearly accessible point of transfer across sectors is at the apex of the system, the degree-granting technical institutes. Furthermore, the looming question of where the private sector--which, if only politically, cannot be significantly disadvantaged by such a policy shift--might fit into a changed system has not been addressed by the reforms discussed or proposed to date. (Note 10)

We argue that this set of policy issues requires an ensemble of less piece-meal policy shifts, i.e., a systemic reform, directed at moving the overall secondary and third-level systems away from the present "pure" type towards a flexible, and ultimately "unified" type. But where are the points at which "reform" will be necessary if such a shift in system-type is to occur? We will conclude this paper by highlighting these necessary points for reform and sketching some of the options that seem available at each point.

Expansion of capacity in the third-level vocational system. The aspiration of the many students who are tracked into the vocational sector of the system but want a form of advanced third-level schooling will need to be satisfied. Expansion of the third-level vocational sector is occurring, but many more places are needed. This need could be satisfied if many of the existing private vocational junior colleges became degree- conferring institutions. (Note 11) The legitimacy of degrees from such upgraded junior colleges could be assured by a transitional certification of specific programs in existing institutions (rather than the institutions themselves) by a national body which would accredit programs within institutions and grant the degrees. Such a gradualist approach to change of many junior colleges would also be a basis for manageable state subsidies for the upgrading costs.

A framework for matriculation to such programs must also be developed, and this framework must be directed at the needs of vocational school graduates and the vocational school sector-- and not be a way by which less successful academic students might enter the advanced vocational system. Such a matriculation framework (which could embrace work experience) could also be a basis for a curricular integration across the vocational sector by serving as a focus for either a more general upgrading of vocational curricula and for the emergence of a clear college-preparatory track within vocational schools.

The private vocational sector. As we noted above, over 60 percent of places in vocational schools are in the private sector. And, as we also found, the costs associated with private schooling, and the lesser quality of private schooling, represent one of the major points of widespread criticism of Taiwan's education policies. Two related options are available to address this problem--although neither, we would argue, is immediately feasible. First, the state could expand the public school component of, particularly, the vocational sector--with the implication that its schools would aggressively (and successfully) compete with the private sector. Second, the state could, as it were, take over all or part of the private sector by either providing operating costs for private schooling or by way of outright purchase of individual schools. However we believe that policies directed at one or another form of take- over of the private sector, which would all involve substantial new state expenditures, are unlikely given the expanding commitments of both central and local governments to increasing social expenditures. (Note 12)

Realistic acknowledgment of the constraints on the state's capacity to support private schooling leads to the possibility that the problem of the private schools' lack of competitiveness with the public school can be addressed not from the point of view of inputs but, rather, by addressing the outputs of the sector. What can make private schools more attractive in the sense that parents can see that their fees are being well spent? This possibility would involve strategies which can improve both the educational quality of the private sector and its articulation with the third-level system.

Vocational schooling requires the continuous renewal of its content and structures in order to respond to changing employment structures and occupational skills. Centrally- controlled and standardized curricula of the kind now mandated by the Ministry of Education (MOE) cannot produce such adaptability but, rather, only serve to limit schools' capacity to keep pace with changing workplaces. A deregulation of private schooling would allow schools to respond more directly to market demands and provide space for schools to develop specialties and reputations for excellence. Such deregulation would involves a shift of focus by the Ministry away from the management of the private sector by process- oriented regulations and towards a monitoring of the outcomes of the school. Additionally, MOE could support a market-oriented development of the private sector by way of funding for, for example, costs of program development, plant and equipment renewal, and the like. Such programs do not, of course, address the issues of equal opportunity and equity which circle pervasively around the present pattern of differential state support for private and public schooling! But in the short and middle run these issues can, in all likelihood, only be addressed by the kind of expansion of the state (and academic) sector currently being initiated in and around Taipei where existing slack middle school capacity is being used to create new state academic secondary schools. (Note 13)

Towards a unified system. Reforms of the kind that we have been outlining can, we believe, address some of the immediate problems of the Taiwan secondary sector. However the long-run problems that are associated with the overall transition of the role of the school in the pathway to adulthood of Taiwan's adolescents remain. We have argued that two socially and educationally differentiated secondary school sectors offer an unstable structure for the provision of schooling in a democratic, egalitarian and increasingly wealthy society. Widely distributed wealth leads to expectations that can only be satisfied by access to higher-status education. The third level will expand its sway over the pathway to adulthood and occupational preparation and, with this expansion, will come ambitions for much greater access college-preparatory forms of schooling.

What kind of policy developments do such possibilities foreshadow? The answer to this question depends in part on the nature of the eventual "target" that is envisioned which, following Raffe (1993), could be a flexible multi-track system or a unified system. However, as Raffe observes, while there has been considerable interest in the idea of unified systems in western societies, there are in fact few examples of such a system in operation. The stratification function of secondary education emerges again and again as proposals become reforms that in most cases constitute one or another form of flexible multi-track system rather than a truly unified system. We predict that the same pattern will emerge in Taiwan: reformers of different stripes will explore the possibility of a unified system, but any reforms that emerge will be in the direction of a more flexible multi-track system. Thus the proposals we sketched above presume the continuation of a multi-track system for the foreseeable future.

What then would be involved if Taiwan's secondary schools were to move more firmly in the direction of a truly flexible multi- track secondary system? Following Raffe (1993; see Table 4 above), the key principles undergirding such a change would center on incrementalism in students' decision-making about their educational futures. This would, in its turn, depend on clear opportunities for transfer between tracks at the second and third level; such opportunities would, in their turn, require the availability of points of transfer between tracks, significant elements of common curricular content between tracks, mechanisms for credit transfer, and narrowed differences between tracks in curriculum, pedagogy, and assessment. Within the present structures of Taiwan's school system, the most obvious immediate constraint on the operationalization of these principles is found in the structures of the Joint College Entrance Examination (JCEE). Vocational students do not have ready access to the JCEE because their curricula do not match the content of the JCEE achievement tests; they cannot enter the mainline college sector because of they cannot participate in the JCEE.

If the JCEE is to remain as the primary mechanism for allocating third-level opportunities to high school graduates as well as the mechanism of curricular articulation and academic selection between the second and third levels, mechanisms by which vocational school students can be given access to the JCEE must be developed. Practically this means that some "new" form of the JCEE will need to be developed which brings into one frame both general and the vocational schools and curricula *but does not, in doing so, submerge the variety and the distinctive missions of the vocational sector*.

The issues which surround such a reform are complex and beyond the scope of this paper. However it is clear that any modification of the examination that retains its character as an academic achievement test rather than an aptitude test will require the development of a framework of a core and options. Such a structure would, in its turn, provide a framework within which vocational schools could develop college-preparatory tracks which could provide access to the third level. At the same time such a development would be a basis for a merger between the curricula of the academic and vocational schools-- for at least some students.

While reform of the JCEE would permit a greater integration of curricula across tracks directed from "above," the plans of the Ministry of Education to extend compulsory education from nine to ten years and the recent proposal to abolish the secondary school entrance examination and reform the university-entrance examination system provide a framework for a curricular integration from "below." Thus we would argue that this extension of the compulsory school should be accompanied by the development of a common 10th grade curriculum across all schools and school type--thus reducing the time (and thus the coverage) associated with the curricula oriented towards the JCEE, and, perhaps, making the difficulties of transfer across tracks less insurmountable than they now are. And were the extension of compulsory schooling accompanied (as it surely must be) with one or another form of a voucher system, the private costs of the private vocational system would be significantly reduced. (Note 14)

Conclusion

While there have been proposals made by Taiwan's "progressive" opinion leaders for the early creation of a publicly-supported comprehensive upper secondary system, we have argued in this paper that this goal is unrealistic in the light of present structures and state policies and priorities. Yet there are major problems within and around upper secondary education in Taiwan which for political, social and educational reasons must be, and will need to be, addressed by Taiwan's educational planners. We have suggested that such planning must be directed to

- the widely understood opportunity costs, in terms of access to the increasingly valued third level, associated with the pure dual-track system of provision of upper secondary education, and;
- the widely perceived inequities in family costs of attending a private school, the most common school type in the vocational upper secondary track where 80% of students are enrolled.

We have argued here that the middle-run "solution" to these problems centers on the conversion of the present multi-track system from a "pure" to a "flexible" form by addressing the points at which barriers inhibit the emergence of a mass college-preparatory larger system from the present élite college-preparatory system. We have further argued that such developments should and must include the private sector. We suggest that policies which will effect a gradual merger across the upper secondary school sectors and across private and state providers will diminish the widespread sense of denied opportunity and/or inequity that trouble the present system.

Notes

1. Net enrollments are calculated as the ratio of the enrollment at a school level of students of specific ages to all people of that age in a national population. Gross enrollment rates are calculated by determining the ratio of students in given grades in a school as a proportion of the population of the appropriate ages. Taiwan's gross upper secondary enrollment rate in 1995 was 91 percent.
2. For the background to the expansion of the vocational education sector, see Li (1995).
3. Although places in the third-level vocational system have rapidly increased in recent years--from 18,000 in 1985 to 51,000 in 1990 and to almost 78,000 in 1995--there are many fewer third-level opportunities for vocational students than for academic high school students (Ministry of Education 1996: 117).

   In 1996 the Ministry of Education (MOE) "required" or "allowed" universities and colleges to admit more high school graduates than in the recent past. It is estimated that 80 percent of academic high school graduates will enter third-level institutions in 1997.
4. In 1990 the per student expenditure in public general secondary schools was NT$51,516 when compared to NT$39,557 for private schools. In 1995 the tuition and fees in public academic school ranged from NT$4040 to NT$4290 while private academic schools charged from NT$13,770 to NT$20,320. The per student expenditure (1990) in public vocational schools was NT$68,624 as compared to NT$44,953 in private vocational schools while tuition and fees in public vocational schools ranged from NT$3500 to 4410 as compared to a range from NT$19,550 to 24,260 in private vocational schools. (Per student expenditures are taken from Chen [1990]; the data on fees in public and private schools was from the Department of Secondary Education and the Department of Vocational and Technical Education of the Ministry of Education. Tuition and fee levels for private schools are prescribed by the Ministry of Education.)
5. Chen (1990) reports that the per student subsidy in public academic schools is NT$46,000-48,000 and for public vocational schools NT$63,000-68,000; she estimates that private academic schools receive a per student subsidy of NT$6593 and vocational schools NT$4111.

The intersection between private and public provision and the organizational form of general and vocational secondary education defines yet another issue in Taiwan's secondary education system. Thus public junior colleges and public vocational schools with their lower fees and higher spending become a preferred sector within the vocational sector while the curricula of these schools makes them (overall) a higher-status third technical (rather than vocational) sector within the overall system. They are schools which confer, moreover, advantages in terms of access to the very limited (in terms of opportunities) technical sub-system of the third-level system. They do not confer ready access to the university system.

The distinction between school types also interacts with social class. Chung (1989) and Yang (1994) report that students of higher social status attend academic high schools while students of lower status attend vocational schools. The sharp difference in SES across school types is obviously problematic from the point of view of equity. Students from families with lower SES pay higher tuition than those of higher SES.

6. For brief discussions of such priorities in the larger Chinese context, see Pepper (1991).
7. Of course, the "public" groups that are salient in determining regime legitimacy will vary depending of the order.
8. Only about four percent these vocational students were successful whereas 51 percent of the academic high school registrants passed the examination.
9. Students from two middle schools were group-interviewed. One school was in a working class community in suburban Taipei County; the other was in an agricultural and fishing community between the middle and south sides of the island. Each school was "typical" in terms of school policies and school size; in each case less than 10 percent of the graduating grade enter a academic high school.

   In the case of the suburban Taipei school, students from three 9th grade classes, two "normal" and one "vocational" were selected for interviews. The parents of the students were interviewed individually, either face-to-face or by phone. In the other middle school students from two classes, one "normal" and one "advanced" were group-interviewed.

   Students and parents from four vocational school were also interviewed: a public commercial school in Taipei City, a private home economics school in Taipei County, a private industry school in Taipei county, and a private nursing school in Taipei City.
10. "The private sector has contributed a lot to the country's schooling. It is not ethical to drive them out of business. Instead the government should be grateful for their contribution and help them financially". (Interview with Chin-Ji Wu, head of the Department of Vocational and Technical Education, Ministry of Education).

    Current discussion of private schooling centers on deregulation of the sector (as opposed to the current tight regulation of the sector by MOE). The implicit goal of such discussion is to encourage greater provision of private schooling.

11. Such a reform has precedents. In 1987 nine (state) normal teachers colleges were upgraded to degree-granting teachers colleges.
12. e.g., a national health care system was implemented in March 1995 and a system of monthly allowances for the elderly has been introduced by the DPP for residents of Taipei City. The national government also proposed social aid for the elderly poor in 1996.
13. While such reforms do threaten existing private vocational schools, the removal of secondary level capacity of junior colleges as such institutions become more clearly third-level institutions would remove capacity from the secondary vocational sector.
14. If the value of such vouchers was set in a way that created more or less parity between the private costs of attending the 10th grade in the public and private sectors, the infusion of funding into the private sector would go a long way to enhancing the quality of private schooling.

References

Chen, L. C. (1993). *Research on public educational expenditures at upper secondary education and the redistribution of incomes*. Taipei: National Science Council. (In Chinese).

Cheng, T. J. (1993). Taiwan in democratic transition. In James E. Morley (Ed.) *Driven by growth: Political change in the Asia-Pacific region* (pp. 193-218). Armonk, NY: M. E. Sharpe.

Chuang, S. Y. (1989). A study of the issue of equal educational opportunity in Taiwan's upper secondary education. Master's thesis, National Kausiung Normal College. (In Chinese).

Chyu, L. H., & Smith, D. C. (1991). Academic secondary education." In D. C. Smith (Ed.), *The Confucian continuum: Educational modernization in Taiwan* (pp. 99-165). New York: Praeger.

Council of Education Reform of the Executive Yuan. (1995). Tracking and education: The meaning of extending the academic high school and the university. Paper for the Fourth Meeting of the Second Seminar on Educational Reform: Examining Tracking and the Examination System. Taipei: Author (In Chinese).

Craig, J. E. (1981). The expansion of education. In D. C. Berliner (Ed.), *Review of Research in Education*, Vol. 9 (pp. 151-213). Washington, DC: American Educational Research Association.

Craig, J. E., & Spear, N. (1982a). Explaining educational expansion: an agenda for historical and comparative research. In M. S. Archer (Ed.), *The sociology of educational expansion: Take-off, growth and inflation in educational systems* (pp. 133-157). London: SAGE.

Craig, J. E. & Spear, N. (1982b). Rational actors, group processes and the development of educational systems. In M. S. Archer (Ed.), *The sociology of educational expansion: Take- off, growth and inflation in educational systems* (pp. 65- 90). London: SAGE.

Cummings, W. K. (1997). Private education in Eastern Asia. In W. K. Cummings & P. G. Altbach, (Eds.), *The challenge of eastern Asian education: Implications for America* (pp. 135- 152). Albany, NY: State University of New York Press.

Deininger, K., & Squire, L. (1996). A new data set measuring income inequality. *World Bank Economic Review*, 10 (3), 565-591.

Department of Education of Taiwan Provincial Government. (1994a). *The 20th survey of the destinations of public and private general high school graduates of the 1992 school year*. Nan-Tou: Author. (In Chinese).

Department of Education of Taiwan Provincial Government. (1994b). *The 20th Survey of the Destinations of Public and Private Vocational School Graduates of the 1992 School Year*. Nan-Tou: Author. (In Chinese).

Department of Education of Taiwan Provincial Government. (1994b). *Educational statistics of Taiwan Province.* Nan-Tou: Author. (In Chinese).

Dreeben, R. (1971). American schooling: patterns and processes of stability and change. In B. Barber & A. Inkeles (Eds.), *Stability and Social Change* (pp. 82-119). Boston: Little Brown.

Englund, T (1996) The public and the text. *Journal of Curriculum Studies*, 28 (1), 1-35.

Epstein, E. H., & Kuo, W.-F. (1991). Higher education. In D. C. Smith (Ed.), *The Confucian continuum: Educational modernization in Taiwan* (pp. 167-220). New York: Praeger.

Gouvias, D. (1998). Comparative issues of selection in Europe: The case of Greece. *Education Policy Analyis Archives*, 6 (4). (Available online at http://olam.ed.asu.edu/epaa/v6n4.html).

Laih, H.-L. (1996). The transition of Taiwan's upper secondary education. Ph.D. dissertation, University of Illinois at Urbana-Champaign, Urbana, IL.

Li, K.T. (1995). *The evolution of policy behind Taiwan's development success*, 2nd edition. Singapore: World Scientific.

Marshall, T. H. (1964). *Class, citizenship and social development*. New

York: Doubleday.

Ministry of Education. (1994). *Report on middle school students' failure to attend school*. Taipei: Author. (In Chinese).

Ministry of Education. (1995). *Education statistics of the Republic of China*. Taipei: Author. (In Chinese).

Ministry of Education. (1997). *Educational statistical indicators of the Republic of China*. Taipei: Author (In Chinese).

Organization for Economic Cooperation and Development (OECD), Centre for Educational Research and Innovation. (1993). *Education at a glance: OECD indicators /Regards sur l'éducation: Les indicateurs de l'OCDE*. Paris: Author.

Pepper, S. (1991). Post-Mao reforms in Chinese education: Can the costs of the past be laid to rest? In I. Epstein (Ed.), *Chinese education: Problems, policies, and prospects* (pp. 1-41). New York: Garland.

Raffe, D. (1993). Multi-track and unified systems of post-compulsory education and "Upper secondary education in Scotland": An analysis of two debates. *British Journal of Educational Studies*, 41 (3): 223-251.

Smith, D. C. (1991). Foundations of modern Chinese education and the Taiwan experience. In D. C. Smith (Ed.), *The Confucian continuum: Educational modernization in Taiwan* (pp. 1-63). New York: Praeger.

Tien, H.-M. (1989). *The great transition: Political and social change in the Republic of China*. Stanford, CA: Hoover Institution Press.

Tien, H.-M. (1992). Transformation of an authoritarian party state: Taiwan's development experience. In T.-J. Cheng & S. Haggard (Eds.) *Political change in Taiwan* (pp. 33- 56). Boulder, CO and London: Lynne Rienner Publishers.

Trow, M. (1960). The second transformation of the American secondary education. *International Journal of Comparative Sociology*, 78 (2), 47-62.

Wu, K. B. (1997). Education policies in Taiwan (China) and Hong Kong. In W. K. Cummings & P. G. Altbach (Eds.), *The challenge of eastern Asian education: Implications for America* (pp. 189-203). Albany, NY: State University of New York Press.

Yang, Y. R. (1994). *Education and national development : Taiwan's experience*. Taipei: Kuei-Kuan Press. (In Chinese).

Young, Y.-R. (1995). School as an epitome of the society: education and social change in Taiwan. In G. A. Postiglione & L. W. On (Eds.), *Social change and educational development: Mainland China, Taiwan* *and Hong Kong* (pp. 120-129). Centre of Asian Studies Occasional Papers and Monographs, No. 115. Hong Kong: Centre of Asian Studies, University of Hong Kong.

Yung, K. C.-S., & Welch, F. G. (1991). Vocational and technical education. In D. C. Smith (Ed.), *The Confucian continuum: Educational modernization in Taiwan* (pp. 221- 275). New York: Praeger.

About the Authors

Hueih-Lirng Laih
National Science Council, Taiwan, ROC

hllaih@nsc.gov.tw

Hueih-Lirng Laih received her doctorate in educational policy studies at the University of Illinois at Urbana-Champaign in 1996 after completing bachelors and masters degree in Chinese literature at the National Taiwan University. She currently has a postdoctoral appointment in the Division of Humanities and Social Science of the National Science Council in Taiwan. She also teaches at the Center for Teacher Training at National Chung-Yang University.

Ian Westbury
University of Illinois at Urbana-Champaign

Phone: (217) 244 5811/244 8286
FAX: (217) 244 4572

westbury@uiuc.edu

Ian Westbury is a professor of curriculum & instruction at the University of Illinois at Urbana-Champaign. He is co- editor of *Science, Curriculum, and Liberal Education: Essays by Joseph J. Schwab* (Chicago: University of Chicago Press, 1978, 1981), *Contemporary Culture and the Idea of General Education* (Chicago: University of Chicago Press, 1988), *Second International Mathematics Study: Vol. 1: International Analysis of Curriculum* (Oxford: Pergamon Press, 1989), *In Search of More Effective Mathematics Education: Examining Data from the IEA Second International Mathematics Study* (Norwood NJ: Ablex Publishing Corp., 1994), and the forthcoming *Teaching as a Reflective Practice: The German Didaktik Tradition* (Mahwah, NJ: Erlbaum). As a comparativist, he has worked with the data from the IEA Second International Mathematics Study and, most recently, on reflective and historical analysis of the German Didaktik tradition.

Westbury is general editor of the *Journal of Curriculum Studies* (JCS) and associate editor of *Revista de Estudios del Curriculum*, a new Spanish journal that is emerging as a collaboration with *JCS*. He is a former Vice-President of Division B of AERA and was a member of the editorial advisory board for the AERA-sponsored *Handbook of Research on Curriculum*.

Department of Curriculum & Instruction
University of Illinois at Urbana-Champaign
341 Armory
505 E. Armory
Champaign, IL 61820

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb of the University of New Hampshire: casey.cobb@unh.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

Greg Camilli
Rutgers University

John Covaleskie
Northern Michigan University

Andrew Coulson
a_coulson@msn.com

Alan Davis
University of Colorado, Denver

Sherman Dorn
University of South Florida

Mark E. Fetler
California Commission on Teacher Credentialing

Richard Garlikov
hmwkhelp@scott.net

Thomas F. Green
Syracuse University

Alison I. Griffith
York University

Arlen Gullickson
Western Michigan University

Ernest R. House
University of Colorado

Aimee Howley
Marshall University

Craig B. Howley
Appalachia Educational Laboratory

William Hunter
University of Calgary

Richard M. Jaeger
University of North Carolina--Greensboro

Daniel Kallós
Umeå University

Benjamin Levin
University of Manitoba

Thomas Mauhs-Pugh
Rocky Mountain College

Dewayne Matthews
Western Interstate Commission for Higher Education

William McInerney
Purdue University

Mary McKeown-Moak
MGT of America (Austin, TX)

Les McLean
University of Toronto

Susan Bobbitt Nolen
University of Washington

Anne L. Pemberton
apembert@pen.k12.va.us

Hugh G. Petrie
SUNY Buffalo

Richard C. Richardson
Arizona State University

Anthony G. Rud Jr.
Purdue University

Dennis Sayers
Ann Leavenworth Center for Accelerated Learning

Jay D. Scribner
University of Texas at Austin

Michael Scriven
scriven@aol.com

Robert E. Stake
University of Illinois--UC

Robert Stonehill
U.S. Department of Education

Robert T. Stout
Arizona State University

archives | abstracts | editors | board | submit | comment | subscribe | search

This article has been retrieved 730 times since October 13, 1998.

Education Policy Analysis Archives

Volume 6 Number 19 **October 13, 1998** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal.
Editor: Gene V Glass Glass@ASU.EDU.
College of Education
Arizona State University,Tempe AZ
85287-2411

Copyright 1998, the EDUCATION POLICY ANALYSIS ARCHIVES.Permission is hereby granted to copy any article provided that EDUCATION POLICY ANALYSIS ARCHIVES is credited and copies are not sold.

The Internet and the Truth about Science: We Gave a Science War But Nobody Came

George Meadows
Mary Washington College

Aimee Howley
Ohio University

Abstract

Even though sophisticated discussion of the nature of scientific claims is taking place in the academy, public school teachers of science and mathematics may harbor naive assumptions about the way that scientific processes function to construct the "truth." Reluctant to change their prior assumptions about science, such teachers may become vulnerable to information technologies (including "low-tech" media such as textbooks and films) that construe science as a collection of facts. An on-line lesson about constructivism provided a forum in which a group of teachers revealed well-established epistemologies seemingly inimical to the principles of conceptual change teaching. Further, the strategies used by the teachers to quell a potentially interesting debate provided preliminary evidence of differences in the motives for communication in virtual, in contrast to real, communities.

665

> "The teaching of mathematics and science is often authoritarian; and this is antithetical not only to the principles of radical/democratic pedagogy but to the principles of science itself. No wonder most Americans can't distinguish between science and pseudoscience: their science teachers have never given them any rational grounds for doing so. . . . Is it then any surprise that 36% of Americans believe in telepathy . . . ?" (Sokal, 1996)

Introduction

Those of us who are interested in science and science education are familiar with controversies involving science and the role of science in society. Recent decades have seen concern over the link between science and the military, threats to the environment posed by new technology, and the implications of advances in biotechnology. Conflicts between science and organized religion, such as the ongoing battles around the topic of evolution, also occur (Gould, 1980; Montagu, 1984; Nelkin, 1982) Whereas the issues raised by these conflicts are important, they do not strike at the foundations of science. Rather, while assuming--sometimes in a naive way--the epistemological claims of science, they question its uses and applications.

Fundamentally different, though, is the battle currently taking place, described by some as the "science wars" (McMillen, 1996). This debate does not focus primarily on the ways that science will be used or on how science threatens certain theological beliefs. Fundamentally more radical, this discussion questions the foundational claim of science, that science can provide an objective view of the physical world.

Advanced by scholars in the social sciences and humanities who study and describe the cultural, social, and political influences on science, this discussion often calls upon arguments from postmodernist philosophy. These arguments concern our ability as human beings to separate knowledge of the world from our personal and social constructions of it. Postmodernists suggest that we each play an important role in constructing our own reality. Given the importance of different social and cultural influences--language being the foremost example--our individual realities cannot be expected to coincide. On this view, scientists have always been and continue to be as vulnerable as the rest of us to the influences of personal experience, culture, and language. Whereas scientific interpretations of the world may be more systematic than non- scientific interpretations, they are not necessarily more true. This argument, of course, challenges the privileged role of science as the sole interpreter of the "real world" (Anderson, 1990).

The recent "Sokal controversy" provided dramatic evidence of the degree to which the "science wars" are now escalating. Physicist Alan Sokal submitted an article questioning the objective basis of science to the journal, *Social Text*, an important postmodern journal (Berkowitz, 1996). His article expressed many of the ideas and views held by postmodernists and carried the additional cachet of being written by a scientist. Shortly after the article was published, Sokal published a second article revealing the first as a hoax. According to Sokal, his aim in perpetrating the hoax was to challenge the academic standards of those scholars who endorse and contribute to postmodern theory. In his view, the editorial board's decision to publish his first article was proof of the slipshod standards and gullibility of scholars in the postmodern camp.

Because this controversy has widened, now encompassing a number of academics from different disciplines and universities (McMillen, 1996), we wondered what meaning it might hold for the classroom science teacher. Academics are raising and debating fundamental questions about the nature and status of science. Should public school science teachers remain unaware of the issues being raised? Should they remain distant from these important discussions?

Recently we had the opportunity to introduce at least some aspects of the discussion to a group of science and mathematics teachers. Whereas we did not plan to discuss the "science wars" per se or to review the Sokal controversy, we did hope to engender discussion of some of the underlying questions: "How do we, as individuals, come to judge what is or is not science?" "What are the boundaries that we establish for our personal beliefs?" "Do these boundaries coincide with those set by science?" It was our intention, through the medium of an on-line course--part of a project with which we were involved--to ask teachers these questions, evoking discussion of foundational issues and relating these issues to the practice of teaching science.

Background

The on-line course was part of the West Virginia K-12 Ruralnet project, an NSF funded initiative whose primary work was to train and assist West Virginia science and mathematics teachers to use the Internet in a variety of ways that enhance classroom instruction.1 Over the forty-two months of its duration, the project worked with approximately 1000 teachers from throughout the state and from every grade level.

In its initial phase, the project provided a two-week summer training session to a group of approximately 40 teachers who were selected to serve as teacher-leaders over the course of the project. These teacher-leaders also participated in two on-line courses, for graduate credit, offered through the two universities involved in the project. Whereas the fall on-line course focused on the practice and development of Internet skills, such as the use of e-mail, listservs, and gopher, the spring course concentrated on the use of Internet resources in the classroom.

Following guidelines set forth in West Virginia's new Science Curriculum Framework, the Ruralnet project advocated a constructivist approach to science teaching. Constructivism not only provides a philosophical framework for the teaching of science, but, as we will discuss below, offers special lessons for the use of the Internet in the classroom. It is this notion of constructivism that lies at the heart of the "science wars" as well. Constructivism raises questions about how our own experiences, ideas, and concepts affect what we come to know through science. It challenges conventions of science instruction that represent science as an absolute and objective picture of the world.

Guided by this approach, we decided that the initial lesson for the course would involve the teacher-leaders in an exercise that implemented constructivist philosophy through conceptual change teaching. Simply stated, conceptual change teaching suggests that learning situations involve the following steps: (1) allowing the learner to state his or her initial concept of a particular phenomenon, (2) engaging in evidence gathering and discourse, debating the merits of different concepts, and (3) restating more adequate concepts (Posner, Strike, Hewson, Gertzog, 1982) This, of course, is an iterative process, continuing as long as time permits. Learners continue to develop their concepts through the process of examination and discourse.

We adopted these steps as the basis for the lesson. We would first provide the teacher-leaders with a topic and ask them to post their initial concept of that topic. The next few weeks would consist of gathering evidence, posting that evidence, engaging in online discourse, and restating concepts. The teacher-leaders would post a final conception and then address several questions regarding how and why they experienced conceptual change.

A good deal of consideration was given to the topic we would discuss. As different phenomena were suggested, we noticed that several criteria were emerging:

- The topic should engender discussion about the nature of science, the scientific method, or what constitutes scientific evidence.
- The topic should not be one where a few experts might dominate the discussion by providing the one "right" answer or explanation.
- The topic should be one about which all teacher-leaders might feel confident in offering opinions.
- The topic should be one to which all teacher-leaders should have had some exposure: we should avoid esoteric, little-known areas of knowledge.
- The topic should be somewhat controversial, but not one in which individuals might place a high degree of value; for example, the topic of creationism might threaten religious beliefs.
- The topic should not be one for which the teacher-leaders would be able to go to a book to find out what they think they should know.
- There should be a good deal of information concerning the topic available on the Internet.
- And perhaps most important, this topic should be compelling enough to engage people in on-line conversation.

The topic we chose was psychic phenomena (i.e., fortune telling, ghosts, channeling, and so on). In addition to meeting our criteria, this topic also was timely: news had just come out concerning the expenditure of millions of dollars by the Defense Department for psychic investigations; a recent broadcast of NOVA, the science-oriented television show, had discussed the evidence for various psychic phenomena; and commercials advertising psychic "readings" were becoming fairly common on television and radio.

We also suspected that the teacher-leaders might have some personal anecdotes or feel comfortable in sharing some "friend-of-a-friend" stories in regard to this topic. Whereas the topic is controversial, we felt that it would not be threatening. We did not believe that it would be linked to value issues, such as religion or politics. It also seemed to be a fairly easy topic to discuss, not requiring technical knowledge or a specialized vocabulary. We suspected that there would be few, if any, authorities on the topic among the teacher-leaders. Additionally, there is a great deal of information available on the Internet in regard to this topic (Sheaffer, 1996). There are numerous sites for skeptics and believers, as well as for the just plain curious.

Furthermore, the topic certainly applies to science, perhaps even challenging conventional wisdom about what might constitute scientific method, reasoning, and evidence. Many of the Internet sites dealing with this topic provide data, discuss research, and "look" scientific (e.g., Princeton Engineering Anomalies Research, on-line), yet the majority of scientists are skeptical of many of the claims made by these investigators (Schick & Vaughn, 1995.) In fact, the scientific appearance of some questionable sites raises a critical issue we had not considered in our initial thoughts on bringing the "science wars" to classroom teachers. If the written word carries power, what kind of power is carried by the animated-graphical-hypertexted-morphed-video-clipped word? The World Wide Web provides a very large audience to just about anyone who can put up an attractive web page. As teachers browse pages (or use search engines to locate sites related to various science topics) how will they be able to judge what is "good" science as opposed to what is "bad" science?

In summary, we felt that the subject of psychic phenomena met our criteria very well. It would provide a good test case for the application of constructivism (and conceptual change teaching) to the use of the Internet in the classroom. Learners would not be given meaning; they would construct meaning through a process of social negotiation. This is where the true value of the Internet in this experiment became obvious. How else could we engage over 40 teachers, from different grade levels and from throughout the state, in social negotiation? What other environment offered such a forum? The participants had common access to a vast amount of information and the ability to communicate almost instantaneously with a relatively large number of peers.

The exercise would also provide the teachers with a framework from which they might view the "science wars." Without some experiential grounding, the claims of the postmodernists seem to be without merit and would most likely be dismissed without consideration. Teachers armed with the experience of seeing how their own constructions are determined and changed would be more likely to gain from theoretical exchanges about the nature of science and its epistemological claims.

Response to the Lesson

We started the lesson with a discussion of its aims, acknowledging explicitly that the topic--psychic abilities--provided a case in point. We indicated in our opening messages that the activities involved in the lesson had more to do with constructivism than with psychic abilities per se. We asked the teachers to suspend their disbelief and to view the lesson as a simulation of conceptual change teaching.

Despite what we thought to be a forthright yet inviting introduction to the lesson, the teachers were not especially receptive. They were neither interested in its constructivist focus nor accepting of its incorporation of psychic abilities as the example of a controversial topic. Several teachers claimed that the lesson was a "joke" or a "waste of time", and a number of them found the topics (both the topic of psychic abilities and the topic of constructivism) "irrelevant", "lacking in interest", and "useless" for them as teachers. This disposition, shared among many of the discussants, may have been responsible for their reluctance to engage with the lesson in the playful, yet serious, manner that we had hoped they would embrace. Despite their reluctance, the teachers did undertake a rudimentary discussion of the topic, which revealed their general stance toward psychic abilities, their strategies of argument, and a surprising but important recontextualization of the issues at stake.

Analysis of the e-mail exchange revealed that, in general, the teachers expressed one of three possible stances toward psychic abilities. Some teachers adopted a stance of uncritical rejection. One teacher's characterization exemplifies this approach: "there is no such thing as psychic ability ... I believe my statement to be undeniable". Another stance embodied uncritical acceptance, characterized by statements such as the following: "although rare ... psychic abilities do exist in certain individuals". Despite the fact that these two stances represent contrasting opinions, neither is critical because neither depends upon nor calls for warrant of any type. Both approaches tend to conflate opinion with true belief, and most of the teachers seemed willing to treat unsupported opinion as sufficient warrant in and of itself. A third approach invoked open-mindedness in dealing with the question of psychic abilities. Some of the teachers who took this approach did so because they did not have a definitive position about the topic-- they spoke of "not closing doors". Others seemed to adopt it because they subscribed definitively to a "scientific" way of thinking, construing science as a method that "always allows for the possibility" of new discoveries. Under this latter construction, the very process of science would require the teachers to take a skeptical rather than a dogmatic stance toward the question.

After making their initial claims about psychic abilities, the teachers provided arguments to elaborate their positions. These arguments tended to be naive, in that they almost always belittled the possible merits of opposing positions. For example, one teacher argued, "I cannot in all seriousness, believe that 'my personal psychic' can tell me what lies ahead for $2.50 for the first minute...." By equating all psychics with "my personal psychic", this teacher challenged the seriousness of any claim that psychic abilities might really exist.

Most of the arguments provided by the teachers subscribed to this general perspective, though there were some interesting variations. A number of teachers chose to "explain away" psychic abilities rather than to give reasons for believing that such abilities are not real. According to one teacher, "a large quantity of so-called 'psychic experiences' are schemes to make money." Others called them "delusions","coincidences", "good guessing," "scams", "hunches", and "our own subconscious controlling our minds". These characterizations, which constituted the most prevalent claims made over the course of the entire discussion, served to distance the teachers from the topic, keeping them somehow immune from it. This strategy was surprising in light of the fact that a few participants did engage the question earnestly and offered some compelling arguments on both sides of the issue. Teachers who distanced themselves from the earnest thread of the discussion tended to marginalize the efforts of those who remained engaged.

One of the arguments, offered by two or three of the teachers who took the discussion seriously, attempted to account for the possibility that psychic abilities might exist. These teachers argued that intuition was part of everyday experience and that psychic abilities might, therefore, involve extraordinary intuitive talent. They also made the claim that the brain had "uncharted reaches" that might house abilities as yet undisclosed.

The most sophisticated arguments offered in the discussion took an inquiring stance and tended to invoke the scientific method as a truth test for the claims made by psychics or by those who believe that psychic abilities exist. Teachers who argued from this vantage seemed to maintain that the burden of scientific proof fell to those making claims about powers that were not within everyone's experience. According to one teacher, "extraordinary claims of any sort require extraordinary proof." Another teacher called for controlled experiments with replicable findings. And another suggested that the scientific community had already reached consensus on the question. Though different, these arguments all spoke to the requirement that such questions be approached both publicly and systematically.

A less sophisticated, but still serious, form of argument relied on personal warrant. This approach was used by teachers arguing on both sides of the question. Several female teachers spoke of "mothers' intuition" as an almost-psychic experience. Others recounted experiences of clairvoyance that could not be explained in conventional terms. And a few teachers used the fact that they had never had psychic experiences or seen demonstrations of psychic abilities as evidence that such experiences and abilities do not exist.

A final class of arguments relied on a fallacy known as "the fallacy of accident". In this case, teachers argued from the general to the particular without attending to the specifics of the particular circumstances. For example, one teacher claimed that if psychic abilities exist then "I do not think some tragedies like the Oklahoma bombing or the Challenger explosion would occur". This reasoning suggests that the existence of tragedies renders impossible the existence of psychic abilities that might predict such tragedies. It doesn't take into account the variety of possible conditions that could mediate the direct connection between any prediction and the actual event or the circumstances that might keep any such prediction from being made, on the one hand, or becoming public knowledge, on the other.

Rather than arguing about the existence of psychic abilities, a few teachers sought to reframe the question in ways that we never anticipated. These teachers contextualized the question within the spiritual rather than the empirical domain and then used Biblical text to warrant their views about it. One teacher wrote: "I do not have any scientific evidence for the existence or non-existence of psychic abilities. However, as a Christian and a believer in the biblical records presented in the Bible, I would have to believe in the existence of psychic abilities." Another teacher, accepting the Biblical claims for the existence of such abilities, cited Leviticus 18:10-12 as a caution against the use of such abilities: "Let no one be found among you ... who practices divination or sorcery ... or who is a medium or spirit or who consults the dead. Anyone who does these things is detestable to the Lord...." As with some of the other argumentative strategies used, this recontextualization of the question assumed a stance that was so definitive that it served to protect teachers from the discussion rather than involving them in it.

This stance, as it was articulated in response to the original version of the question as well as to the recontextualized version, managed to render as unarguable a topic that the lesson identified as prototypically arguable. It clearly transformed the nature of a dialogue that was supposed to constitute and exemplify "constructivist" teaching and learning. It is not clear to us whether or not this transformation was intended by the teachers as a way to defeat the premises of the lesson. But it does seem apparent that their assumptions, dispositions, and modes of arguing actually had this effect.

At the end of the lesson, the teachers were so distressed by the discussion that they were unwilling to respond to our efforts to debrief. We had hoped that the dialogue about psychic abilities would provide a shared experience from which we might all examine the practice of conceptual change teaching. The most vocal of the teachers, however, made clear their displeasure with constructivism, identifying it as an esoteric theory with little practical import for public school classrooms. If there were teachers in the group who were supportive of constructivism, the tenor of the discussion was sufficiently hostile to insure their silence.

Using the Internet to Assist Meaning-Making

As the result of this less-than-successful lesson, we learned a number of things about the nature of discourse and the ways that Internet use can interfere with it. First, we received an important reminder about the strength of prior assumptions. Working from a constructivist vantage, this was no surprise in a theoretical sense. But we did not anticipate the important difference between specific naive assumptions and well-formed, internally-consistent sets of assumptions based on alternate world views. In short, we found that, among many of the teachers, prior constructions of reality (and of science and also of discourse) were not sufficiently piecemeal to admit conceptual change. Rather, the coherence of their views--the religious beliefs of some of the teachers as well as the pedagogical beliefs of most of them--made them resistant to the cognitive dissonance that the lesson attempted to provoke. Put another way, the teachers' prior assumptions were sufficiently elaborate and functional as to make assimilation relatively easy and accommodation almost impossible. Thus the social negotiation that we had hoped to stimulate was rejected because it stood outside of the belief systems of the teachers. In a very real sense, discourse of this type did not exist for them.

Obviously, the clash of belief systems characterizes all discourse, not just the discussions that the Internet permits to take place. But, because of their nature, virtual discussions in virtual communities may pose particular dangers to discourse in general. Unlike physical communities, virtual ones share no common ground in the very literal sense.2 Grounded in other shared purpose (e.g., the cultivation of a neighborhood that belongs to everyone), physical communities allow multiple perspectives to exist side by side, interacting and having cross-influences over long periods of time. Members of physical communities have some stake in maintaining a peaceful way of life, and they offer shared activity as a solace for the losses encountered in clashes over belief. But the stakes in virtual discussions are not very high, and the requirements for mannerliness are, therefore, formal rather than implicit. Furthermore, in the absence of the physical encounter, virtual discussions reduce all discourse to mere words. The relationship between words and a way of life is lost in this forum. This loss is important because it reinforces the already rampant alienation and narcissism of our late twentieth century society--supporting the logically insupportable argument that all beliefs have an equal claim to truth, that all values are equally good, and that personal inclination is the final arbiter of both truth and merit.

Added to this disturbing circumstance are other features of virtual life that we observed to become animated in the lesson on psychic phenomena. Important among these features was the tendency of the Internet to disable efforts to distinguish between reputable and disreputable sources of information (see e.g., Burbules, 1996). Almost anyone can have a web page, and almost anyone can post a message to a discussion list. Moreover, these artifacts can take the form of very credible-looking products. At the same time such products need not contribute anything of substance; they can mislead unintentionally or intentionally.

Some commentators suggest that this feature democratizes discourse, and it may indeed have this effect; but the caveats necessary to accommodate this type of democratization may be so intrusive as to inoculate all discussion from credibility. Without having traditional sources of intellectual authority to rely on, one might as well invent reality capriciously. An alternative, of course, is to hope that everyone will become sufficiently knowledgeable, critical, and sophisticated so as to be able to distinguish routinely among the multiplicity of competing truth claims. From our experiences with this lesson, however, we suspect that a third strategy may have wide currency: In the face of multiple, incompatible, and seductive truth claims, people may very well do what the teachers in our group did--retreat more deeply into their previously held belief systems, shield these systems from intellectual challenges, and refuse to entertain serious argument across assumptions.

Implications for Science Teaching

The approach taken by the teacher-leaders with whom we worked effectively removed them from discussions about the nature of science and scientific claims (cf. Pomeroy, 1993). These discussions, however, may be critical to informed practice of science education since they implicate both the method and the findings of science. Scientists--no matter what their take on the "science wars"--avoid the naive claim that science establishes an infallible canon of natural law. Notably, proponents on either side of the debate promote more subtle and sophisticated views of science than our teacher-leaders were willing to entertain. This circumstance is more troubling than the "science wars" themselves, which, after all, entail thoughtful, dynamic regard for an important realm of human inquiry.

At a time when science teachers need to be increasingly careful in sifting through vast arrays of information, reliance on established "fact" seems to be a most unfortunate anachronism. Encouraged to accept constructivist aproaches, science and math teachers still cling to traditional rote and text-based methods (Besvinick, 1988; Gess-Newsome & Lederman, 1991; Stigler & Hiebert, 1997). Although structural constraints clearly do keep science and math teachers from changing their instructional methods to incorporate constructivist practices (e.g., Keiser, & Lambdin, 1996), our investigation suggests that their prior beliefs about science teaching and about the nature of science itself may constitute another--possibly more formidable--impediment to change.

Notes

1. The authors wish to acknowledge the National Science Foundation's support for the West Virginia K-12 Ruralnet project (NSF 95-50017) and the research conducted in conjunction with that project.
2. We appreciate and agree with the comments of a reviewer of the article who reframed our distinction between real and virtual communities more broadly to encompass the distinction between real communities and arbitrary groupings of people (e.g., in

classrooms, in the work place, on the freeway).

References

Anderson, W.T. (1990). *Reality isn't what it used to be.* San Francisco, CA: Harper San Francisco Press.

Berkowitz, P. (1996, July 1). Science fiction: Postmodernism exposed. *The New Republic* [On-line]. Available: http://www.enews.com/magazines/tnr/archive/07/berkowitz070196.html

Besvinick, S.L. (1988). Twenty years later: Reviving the reforms of the '60s. *Educational Leadership*, 46(1), 52.

Burbules, N. (1996). Technology and changing educational communities. *Educational Foundations*, 10(4), 21-32.

Gess-Newsome, J., & Lederman, N.G. (1991, April). Preservice biology teachers's subject matter structures and their relationship to the act of teaching. Paper presented at the annual meeting of the National Association for Research in Science Teaching, Lake Geneva, WI. (ERIC Document Reproduction Service No. ED 331)720

Gould, S.J. (1980). *The panda's thumb.* New York: Norton Press.

Keiser, J.M., & Lambdin, D.V. (1996). The Clock is ticking: Time constraint issues in mathematics teaching reform. *Journal of Educational Research*, 90(1), 23-30.

McMillen, L. (1996, June 28). The science wars. *The Chronicle of Higher Education*, pp. A8-A9, A13.

Montagu, A. (1984). *Science and creationism.* New York: Oxford University Press.

Nelkin, D. (1982). *The creation controversy.* Boston, MA: Beacon Press.

Pomeroy, D. (1993). Implications of teachers' beliefs about the nature of science: Comparison of the beliefs of scientists, secondary science teachers, and elementary teachers. *Science Education*, 77(3), 261-278.

Princeton Engineering Anomalies Research (PEAR). [On-line]. Available: http://www.princeton.edu/~rdnelson/pear.html

Posner, G.J., Strike, K.A., Hewson, P.W., & Gertzog, W.A. (1982). Accommodation of a science conception: Toward a theory of conceptual change. *Science Education*, 66(2), 211-227.

Schick, T. & Vaughn, L. (1995). *How to think about weird things: Critical thinking for a new age.* Mountain View, CA: Mayfield Publishing Company.

Sheaffer, R. (1996, May/June). The weird world web. *The Skeptical Inquirer*, 20(3), pp. 17, 54.

Sokal, A.D. (1996). Alan Sokal's home page. Available: http://www.nyu.edu/gsas/dept/physics/faculty/sokal

Stigler, J.W., & Hiebert, J. (1997). Understanding and improving classroom mathematics instruction. *Phi Delta Kappan*, 79(1), 14-21.

About the Authors

George Meadows
Department of Education
Mary Washington College

Phone: 540-654-1351
E-mail:gmeadow@mwc.edu **George Meadows** is an Assistant Professor in the Education Department at Mary Washington College. He received his Ed.D from West Virginia University, working in the area of science education and technology. Current interests include the use of conceptual change methods in teaching multiculturalism and the applications of art in science education.

Aimee Howley Profeesor
College of Education
Ohio University

Email: howley@oak.cats.ohiou.edu

Aimee Howley is Professor in the Educational Studies Department at Ohio University. Teaching primarily in the Educational Administration program, her research examines critically the theory and rhetoric that inform educational practice in the US. She is currently working on an analysis of developmentalism as an ideology, focusing particularly on it contradictory influence on US pedagogies.

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

John Covaleskie
Northern Michigan University

Alan Davis
University of Colorado, Denver

Mark E. Fetler
California Commission on Teacher Credentialing

Thomas F. Green
Syracuse University

Arlen Gullickson
Western Michigan University

Aimee Howley
Ohio University

William Hunter
University of Calgary

Daniel Kallós
Umeå University

Thomas Mauhs-Pugh
Green Mountain College

William McInerney
Purdue University

Les McLean
University of Toronto

Anne L. Pemberton
apembert@pen.k12.va.us

Richard C. Richardson
Arizona State University

Dennis Sayers
Ann Leavenworth Center
for Accelerated Learning

Michael Scriven
scriven@aol.com

Robert Stonehill
U.S. Department of Education

Greg Camilli
Rutgers University

Andrew Coulson
a_coulson@msn.com

Sherman Dorn
University of South Florida

Richard Garlikov
hmwkhelp@scott.net

Alison I. Griffith
York University

Ernest R. House
University of Colorado

Craig B. Howley
Appalachia Educational Laboratory

Richard M. Jaeger
University of North Carolina--Greensboro

Benjamin Levin
University of Manitoba

Dewayne Matthews
Western Interstate Commission for Higher Education

Mary McKeown-Moak
MGT of America (Austin, TX)

Susan Bobbitt Nolen
University of Washington

Hugh G. Petrie
SUNY Buffalo

Anthony G. Rud Jr.
Purdue University

Jay D. Scribner
University of Texas at Austin

Robert E. Stake
University of Illinois--UC

Robert T. Stout
Arizona State University

---

archives | abstracts | editors | board | submit | comment | subscribe | search

EPAA
• enter the archives • browse the abstracts • the editors • the edit board
• submit article • submit commentary • search • subscribe
volume: • 1 • 2 • 3 • 4 • 5 • 6 • 7

This article has been retrieved 1235 times since November 3, 1998

Education Policy Analysis Archives

Volume 6 Number 20 November 3, 1998 ISSN 1068-2341

A peer-reviewed scholarly electronic journal.
Editor: Gene V Glass Glass@ASU.EDU.
College of Education
Arizona State University,Tempe AZ
85287-2411

Copyright 1998, the EDUCATION POLICY ANALYSIS ARCHIVES.Permission is hereby granted to copy any article provided that EDUCATION POLICY ANALYSIS ARCHIVES is credited and copies are not sold.

Critical Evaluation for Education Reform

Gisele A. Waters
Auburn University

Abstract

The school reform movement has done little to provide an accurate analysis of the production of inequality or the reproduction of social injustice in the public schools or the larger social order. The ideology that influences this movement has often prevented the realization of any notion of an egalitarian ideal, the elimination of inequality, or the improvement of those who are least well-off. I ask educators and evaluators of education reform efforts to reconsider critically their roles in social science research, to reclaim the battleground of public school reform by focusing on the democratic purpose of public schooling, and the institutional problems in educational programs and practice that often inhibit action toward this ideal. The first part of this article includes an extensive argument explaining the "why" of critical evaluation. The theoretical literature on inquiry in science and social science, the ideology of critical theory, critical social psychology, and Freirean pedagogy are consulted as additional tools for augmenting the practice, policies, and responsibilities of evaluators in education. I review three contemporary perspectives of evaluation in order to begin rethinking the purposes and functions that evaluation serves in education. It also demonstrates how mainstream and contemporary evaluations can be used to serve a particular set of social and political values. The second part of this article begins a preliminary journey toward describing the "how" of critical evaluation. Critical evaluators can fight for social justice by combining the merit criteria of state and federal public education law, and the methods of an adversary oriented evaluation in order to transform educational environments that serve the future potentials of all children. Therefore education involves the practice of freedom, the means by which men and women deal critically and creatively with reality and discover how to participate in the transformation of their world (Freire, 1985).

The Argument for Critical Evaluation of Education Reform

Part I: The "Why" of Critical Evaluation

Schools are inextricably linked to the communities they serve through social, political, economic, and cultural interests. To better comprehend public education, the socio-cultural, political, and hierarchical relationships that transpire within the school as well as within the community must be linked to the broader political and economic issues of society at large (Ogbu, & Matute-Bianchi, in press). To begin to realize the possibility for reforming public education, and to begin fighting for social justice in education, especially for those children who are disadvantaged, we must first re-examine the historical nature of the problems of education and the communities in which these schools exist (Noll, 1997).

Education Reform

School reform reports in the 1980s and early 1990s served to spotlight the nature and the function of public schooling and attempted to delineate a specific relationship between broader social, economic, political, and cultural interests. Kretovics, Farber, and Armaline (1991) state that many of the reports argued in favor of a wide range of additive reforms such as increased testing, more homework, a longer school year, a longer school day, and the internalization of an extensive list of cultural facts. Others suggested a plethora of technical solutions for the challenges facing public schools, ranging from the addition or reduction of certain educational requirements for teachers, to the addition of specific course requirements to the public school curriculum (Noll, 1997). For the most part, the reform efforts have been driven by what Barth (1986) called perseverence of list logic and what Giroux (1988) has called the ideology of the quick fix.

Berliner and Biddle (1997) suggest that there are data that illuminate the untenable assumptions on which much of the school reform movement is based. Similar to Kozol (1991) they also suggest that much of the school reform movement is wrongheaded, and that some of these reform efforts are thinly disguised elitist attempts to get rid of public education, to protect the privilege such individuals have already bestowed upon their children. Coming from a radical social theorist perspective, these claims might be completely dismissed; but Berliner is identified by the scientific community as a respected, traditional educational psychologist. Berliner and Biddle also draw on the work of the systems analysis department of a national laboratory that was under authorization of the Reagan and Bush administrations.

Berliner and Biddle's book, *The Manufactured Crisis* (1997), was partly written to help dispel some of the myths and distorted evidence that federal politicians were creating about public education. Berliner and Biddle examine the real problems of education that have too often been masked. These authors add credence to the notion that some proposals for educational reform reflect only the personal experiences or prejudices of legislators, and that some are based on misunderstandings about schools and the problems of education. For these reasons, many programs intended to "improve" our schools turn out to have little detectable effect or, worse, end up creating serious problems for educators and students.

For more than 20 years, a variety of educational and social theorists have presented compelling arguments that illustrated the reproduction of social, economic, political, and cultural inequalities through the organization and structure of the schooling process (Kretovics & Nussel, 1994). They summarize very succinctly some of the past characters that entered the debate on controversial issues in education and social theory. Kretovics and Nussel state that educators from diverse cultural and ideological backgrounds (i.e., Coleman, Anyon, Giroux, Arnot, Clark, McRobbie, Illich, Bowles, Gintis, and Apple), drawing on the earlier works of other scholars (i.e., Dewey & Baldwin) have pointed to the political and ideological nature of schooling and the ways in which schools often under serve the nonmajority students and, through hegemonic practices, reproduce the status quo. The arguments have often been ignored, dismissed, or co-opted resulting in a blaming of the victims of educational inequalities. As a result, schools in general, are blamed for the broader social and economic problems that inform and structure their existence.

Through the recent school reform movement, the problems of American education and the general purposes of public schooling have been systematically removed from the terrain of public debate (Kretovics & Nussel, 1994). Kretovics and Nussel go on to say that "most of the widely publicized school reform efforts have created an educational climate antithetical to moral referents such as social justice and initiatives of equity that are valued in a democratic society" (1994, p.4). Giroux (1983) and Freire (1985) believe that the schooling process is always structured on the norms and values that embody specific social, political, economic, cultural, and ideologic interests.

Therefore as educational leaders and as evaluators of educational programs we should also critically examine the context, both present and past, in which education takes place to illuminate both the problems and the possibilities for change in the future. Evaluators of educational reform must have a vision for change and the ability and conviction to act on that vision. Evaluators are inevitably linked to one of the most crucial of social processes, education, and must develop a framework that takes seriously issues of power, democracy, inequality, as well as educational structures and practice, in the difficult process of reforming American public education (Shapiro & Purpel, 1998).

Universities and other institutions of education and education evaluation should adopt policies and practice that link their educational contributions closely with improved schooling for America's young or surrender their franchise (Maruyama & Deno, 1992). In the 1960's evaluators called in to assess effects of social reforms, especially Head Start programs, focused on technical outcomes, determined little significant changes, and continued the debate over methodology, the purposes, and the audiences for evaluation. Critical evaluation should question its purposes for conducting evaluations within the context of education and the reform efforts designed to restructure and transform education environments. &

The Conceptualization of Educational Evaluation as Practical Educational Research

Many attempts have been made in recent years to clarify the meaning of evaluation and expose the distinction between evaluation and other related concepts such as measurement or research. The literature contains many approaches regarding the conceptualization of evaluation and the determination of its countenance in education (Nevo, 1986). According to Nevo (1986), many of these approaches have been unduly referred to as "models" (for example, the CIPP Model, the Discrepancy Model, the Responsive Model, or the Goal-Free Model) in spite of the fact that none of them includes a sufficient degree of complexity and completeness that might be suggested by the term "model." For the benefit of those of us who lost their way among the various models and approaches, I simply suggest taking a holistic approach to considering educational evaluation as an extended arm of practical educational research.

Education is a field like medicine in that its name simultaneously refers to a practice and to a field of disciplinary inquiry (Scriven, 1986). Scriven stated that the paradigm of research in the area of the philosophy of education, to take one example, is surely the paradigm of philosophical research in any area. But that leaves open the area of research that we normally think of as the domain of scientific research in medicine or education. Traditionally, we have tended to suppose that in this area of medical or educational research the correct model is that of the related sciences. That is, for example, educational research has modeled itself on social science research. Similarly educational evaluation has modeled itself as an offspring of educational research. In medical research that approach has brought some problems because it seems to lead to results that conflict with the practical wisdom of physicians and the economic realities of the patients. The same can be seen in education with the refined development of IQ tests, norm- referenced testing, and token economies for classroom management.

Scriven (1986) wrote that the conventional "scientific paradigm" way of dealing with these type of problems is not the business of science, they are value issues, and must be sorted out by the citizenry. Instead, he proposes a paradigm for practical educational research which subsumes educational evaluation, and which includes ethics, political feasibility, a set of practical alternatives, and an overall practical significance. Educational research is not, as he is suggesting, to be defined as all research that in any way involves the concepts related to education, because that's too broad (it includes learning theory), but as research that contributes to the facilitation of education, just as medical research should not be defined as all research that involves concepts related to medicine, since that brings in all physiological research, but simply as research contributing to health.

The research on classroom teaching, educational programming, school management, and classroom achievement have mostly been designed on the "quest for knowledge" model (traditional scientific) rather than on the "improvement of practice" model. Scriven's main point for educational evaluation stresses the acknowledgment that evaluation research in schools can be a far more complex business than just a quest for knowledge, a quest for classification, explanation, generalization, causation, and/or prediction.

In reviewing some of the theoretical literature of inquiry in science and the social sciences, it is hard to avoid the impression that there is a reluctance to confront the issues of power, democracy, inequality, ethics, politics, and pragmatics in educational research, evaluation, and in mainstream social science. Scriven proposed that one cannot reconcile the widespread support for the doctrine of a value-free social science with the continued, inescapable practice of evaluation by social scientists, of the work and worth of students, peers, and selves, except by invoking a kind of phobia which makes them blind to the contradiction between their doctrine and their practice. This phobia, Scriven called "value phobia," has blocked us for nearly a century from addressing explicitly the methodology of evaluation and the systematic evaluation of our own practices in social science research (1986, p.62). With this in mind, I explain how the theoretical literature of inquiry in science and the social sciences can contribute to justifying the inclusion of such values as social justice within an expanded framework of critical evaluation of education reform.

Consideration of Inquiry in Science and the Social Sciences

Social Justice and the Distribution of Education

Considerations of social justice are applied in the distribution of virtually every social good. This is so much the case that, in the eyes of some, social justice simply has to be proclaimed (for example in political programs) to henceforth characterize the relations between people. In educational policy, arguments derived from social justice played a role even before World War II and were fought over by political parties, teachers' unions, left-wing intellectuals, and "pedagogical entrepreneurs" (Wesselingh, 1997). For them, the phenomenon of unequal participation was indeed a social problem, a phenomenon of social injustice. It does not take much effort to see that predominantly economic considerations have prompted the rapid expansion of equal- opportunities research. Opinions about the just provision of educational opportunities combined with economic need, have given the impetus to this research (Wesselingh, 1997).

Indeed, the Fall 1998 edition of Educational Evaluation and Policy Analysis includes one of the latest studies completed using Hierarchical Linear Modeling (HLM) on the relationship between students' opportunity to learn (OTL) and their science achievement. In this study, Jia Wang concludes that content exposure was the most significant predictor of students' written test scores, and the quality of instructional delivery was the most significant predictor of the hands-on test scores. In support for these types of conclusions, Berliner and Biddle (1997) clearly argue that opportunity to learn is the most significant predictor of academic achievement. These authors would be content to know that "scientific methods" such as HLM techniques are pushing the analysis of OTL variables at two level of instructional processes: the classroom level and the student level.

On the cusp of a new millennium, we are searching for answers not in the homes, economic backgrounds, and individual disadvantages of our students of public education, it would seem that we are finally beginning to look at the quality of instruction variables that exist in schooling processes instead of "blaming the victim." Can we begin to ask why and how our school systems are failing our children, instead of why and how these children are failing our school systems" If schools are to be held accountable for the equitable delivery of educational opportunities and if social justice is to take place within the halls of academic opportunity, the core of the education performance indicator systems should include school and classroom information.

According to Winfield (1993), there are two main reasons for obtaining OTL information. First and foremost, teacher and school factors need to be taken into consideration in explaining students' achievement. Teacher and school variables directly and indirectly influence student learning and student performance. Second, the new performance-based assessments make the collection of OTL information crucial (National Council on Education Standards and Testing (NCEST), 1992). The performance-based assessments require higher order thinking skills. This may put students from low socio-economic status groups at a disadvantage. Studies have shown that minorities, especially African American and Hispanics, are more likely to be put into classroom with less learning opportunity even when ability is taken into account (Gross, 1993). If future research on achievement continues to disregard OTL variables, the achievement gap between majority and minority will continue to increase and a lack of educational opportunity will continue to expand (Arreaga-Mayer & Greenwood, 1986; Madaus, West, Harmon, Lomax, & Viator, 1992).

> Education as a good to be distributed gets the character of a good that provides access to other goods. The key power of schooling is based on the fact that education serves as a criterion for the distribution of all kinds of other material and immaterial goods. The consequence of this development is an instrumentalization of education. It evolves into an outstanding example of an instrument of mobility in a society where now qualification and rapidly growing demands for qualification create the space for moving up and, to a lesser extent, moving down the social ladder. This is at least the idea; the question of how education actually performs or is able to perform its role as a social agency of distribution for various social groups is of course not answered (Vervoort, 1975, p. 104).

For various groups this question still challenges our daily lives as critical evaluators, leaders and researchers of social justice in education.

As generally acknowledged since the traditional bourgeois ideas of the Enlightenment, the only valid criterion for determining who deserves which education is achievement. Achievement as a criterion for selection stems from egalitarian principles and is generally accepted in education as a just criterion. By now we know that this distribution model has let to serious forms of social inequality. The assumption that in schools everybody has equal opportunities to perform and thereby has a fair chance to take part in the subsequent competition on the labor market, has proven to be a misconception (Wesselingh, 1997). Education thus functions as an instrument for the reproduction of social inequality and thereby reflects the irony of a principle derived from egalitarian Enlightenment philosophy.

Social Justice and Education

Walzer's *Spheres of Justice*, published in 1983, can be seen as a reaction to John Rawls's *A Theory of Justice*, published in 1971. Walzer's objective was to provide an interpretation of what we contemporary Americans see as the essence of such concepts as equality and justice. His book makes clear that a discourse on the selection criteria for such an important social good as education is now needed more than ever. Reflection on this topic should not be left to politicians and policy makers for in that case considerations outside the sphere of justice will tend to dominate. Educational scientists, sociologists of education, and educational evaluators in particular, should definitely be more concerned with issues of social justice in education. Social justice is one of the most important values that we should hope to secure in critical evaluation studies of educational reform.

One of the goals of this article, besides arguing that critical evaluation is needed in order to begin fighting for social justice in education, is to recommend an open and purposeful discourse about social justice in the reform of American public education between the "public intellectuals" (Giroux, 1997, p.263), otherwise known here as the social scientists, educational researchers, evaluators, and practitioners, a discourse about social justice in the reform of American public education. The participation in discourse that values a moral imperative and a political commitment to social justice in the evaluation of education reform is crucial to understanding the ideology of a critical evaluator.

Reaching Beyond the Incommensurable Perspectives

When it comes to dealing with such issues as social justice in education in a way that recognizes its moral complexity and political nature, the social sciences have "incommensurable perspectives" based in various traditions which have had different ideas about the individual and his/her society (Wertsch, 1998). These views have been updated but often at the cost of further fragmentation in the social sciences. The work of the new "public intellectuals" is to translate and connect, the ideologies and contributions of Aristotle's Realism, Plato's Idealism, Comtean Positivism, the Vienna Circle of Scholars and their Logical Positivism, Constructivism, Postmodernism, Critical Social Theory, and Feminist Theory.

The immeasurable challenge of the future is to look through diplomatic eyes without the "terministic screens" (Burke, 1966) of our specializations and disciplines that impair our vision. We could begin to address the phenomenon of public schooling, its reform, and its evaluation within a politically honest analysis. By refocusing our individual and collective powers into a moral and political analysis, critical evaluation of education reform in the next century can begin to regain the democratic imperatives or possibilities of public institutions. Exercising this moral and political "judgement" in evaluation of education reform, as a social responsibility in public practice, requires instrumental courage and conscience.

Analytic Primacy

Also this article aspires to begin a discourse beyond what politicians, educators, and philosophers have debated for centuries, the extent to which education should develop the individual or serve the needs of the state and society. The fact that this debate seems to go on and on with no principled resolution in sight suggests that deeper issues may be at stake. Namely, it suggests that academic dispute over what has "analytic primacy" (Wertsch, 1998, p.9), the individual or society, may reflect an underlying debate, a debate that cannot be resolved through rational argument. I am recommending that evaluators of education reform lift the blinders of methodological habit, move beyond their rational arguments, and discover how their own morals and politics are partly reflected in their professional decisions. With this in mind we live in times of increasing uncertainty as to how to reform public education. Part of the success of education reform will depend on those who have the power to affect social change, who have control over the knowledge base, who judge the worth and merit of educational programs, and what kinds of morals and politics are profoundly ingrained within their minds, spirits, and hearts.

Social Justice and Public Practice

For the most part, educational research and evaluation have remained both moral and political innocents in theory, practice, and policy. Part of this political innocence is derived from self reproducing ideologies and scientific paradigms that have explicitly or implicitly neglected moral and political issues. The conception of social justice, as considered here, is not a privilege for some (meritocractic) but rather a birthright for all (democratic) (Sirotnik, 1990). The value of social justice forms the foundation for working towards the restoration of a moral and political theory in the evaluation of public education reform, as part of a social responsibility in public practice, and as a part of confronting the moral and political purposes of social inquiry and research.

The contributions of Wertsch, (1998), Giroux (1997), Prilleltensky (1994), Tsoi Hoshmand, (1994), Howard (1985), Kohlberg (1984), Rawls (1971), Habermas (1971), and Kuhn (1970) are offered as significant commissions to support the reconsideration of our individual and professional decisions in education, by deliberating on our own morals and politics. Reflections and deliberations on our own values, beliefs, passions, and the reasoning for our professional decisions are mostly done outside of the confines of our "professional lives." Thus we are left with the interesting and paradoxical conclusion that what "ought" to be the most central in the evaluation of our schooling of American children, the moral and political reasoning, becomes inevitably peripheral to our public practice (Miller & Safer, 1993). In terms of articulating in-depth moral and political positions related to evaluation in educational reform, these considerations and decisions are vital to building and transforming schools that are struggling to achieve democratic ideals.

Between the Potential and the Present

Issues such as equality, democracy, race, gender, class, and poverty are certainly integrated through variable means into the contemporary scholarship of educational psychology, research, and evaluation. However, these issues and their historical, political, moral, and economic meanings are rarely discussed in a comfortable forum naturally or agreeably in the impregnable halls of academia. Therefore, the silent space between the potential in education and the present crisis in public education is successfully and safely insulated decade after decade. As a result, inquiry and discourse between "public intellectuals" remain fixed in a non- political environment without values, beliefs, and passions. This environment within an "ideology of neutrality" became internalized in the consciousness of most researchers following the establishment of the modern university. The links between the political agendas and research were, and often remain, blurred by the legitimating function of social and educational research. This can be seen in many educational evaluation studies that accept the objectives of pedagogical programs and are organized to "explain" how the objectives were reached.

Redefinition of Identity and Purpose

No problem is more difficult and complex in the social sciences than that of determining how morals and political values are embedded within the research methodologies that we employ and the "academic" decisions that we make (Cronbach & Associates, 1980; Hamnett, Kumar, Porter, & Singh, 1984; Fetterman, 1988; and Sirotnik, 1990; Maruyama & Deno, 1992). That morals and political values should exist in research is no longer denied (Warren, 1963; Fetterman, 1981; Freire, 1985; Apple & Beyer, 1988; Habermas, 1990; Prilleltensky, 1994; Giroux, 1997; Kanpol, 1997; Wertsch, 1998). In terms of educational evaluation, the ideas found in this article, reconfirm the conviction made by Sirotnik (1990) that the practice of evaluation is part of the political authority structure of society, and that evaluation as an aid to public decision making entails conceptions of democracy and social justice, even when these conceptions are not immediately apparent.

House (1993) wrote that evaluation receives its authority not only from its presumed "scientific method" but also from government endorsement itself. Within the analysis of evaluation in advanced capitalist societies, House reviewed how governments face serious problems in governing such a multicultural "amorphous mass of people" (1993, p.vii) and how evaluation is both political and scientific authority applied to practical decision and actions, particularly public decisions and actions. He went on to explain how governments are capable of making decisions based on their own political authority, but that it is easier to govern based on voluntary acceptance by the populace attained through scientific persuasion, particularly when the populace is pluralistic and increasingly non-traditional. In addition, House expanded the notion of political and scientific authority by redefining formal evaluation as a new from of cultural authority. Cultural authority can be manifested in the probability that descriptions of reality and judgements of value will prevail as valid, an increasingly difficult accomplishment in societies with disparate value systems (House, 1993). Current literature in evaluation confirms that evaluation as a social activity is becoming increasingly self-conscious about its own identity and purpose in the larger social order (Cronbach & Associates, 1980; Guba & Lincoln, 1989; Rossi & Freeman, 1993; Patton, 1994; Scriven & Kramer, 1995; Chelimsky & Shadish, 1997).

Critical Evaluation

Critical evaluation of education reform involves the practice of completing empirical, historical, public and social work by employing explicit theories of justice (House, 1976,1980) that require serious commitment, persistence, courage, conscience, and conviction in order to restructure and transform education environments. Hence, as a social practice, evaluation involves an inescapable ethic of public and social responsibility that extends well beyond the immediate clientele by focusing on the democratic purpose of schooling. Social justice in evaluation, then, concerns the manner in which various interests are served. Critical evaluation should serve the interests not only of stakeholders, sponsors, or the reformers, but of the larger society and of various groups within society, particularly those most affected by the educational programs under review. One of the aims of this article is to reiterate that institutions of higher education must be seen as deeply moral and political spaces in which evaluators, indeed intellectuals, assert themselves not merely as professional academics but as citizens, whose knowledge and actions presuppose specific visions of public life, community, and moral accountability (Giroux, 1997).

A Political Theory

I propose here that critical evaluation represents a kind of political theory that integrates explicitly the value of social justice into the practices, policies, and responsibilities of evaluation of educational reform. Moreover, the political theory of critical evaluation can be defined as the implicit and explicit social and professional ethics of evaluation, and the moral and political consequences of these ethics, which could reconstruct and reconsider the power relations in academia and public education. One of the reasons to begin a journey into a critical political perspective in educational evaluation is to arrive at an account, a kind of "translation at the crossroads" (Wertsch, 1998, p.7), that would make it possible to link, but not reduce, one perspective of "science" to another. Another reason is to begin addressing explicitly the methodology of evaluation and systematically evaluate our own practices in social science research (Scriven, 1986).

The task is to reflect, to discourse, and to collaborate with each other, between and within disciplines, to dialogue about the human condition, especially the conditions of inequalities that public institutions perpetuate in our democratic society. In order to talk and listen to one another about social justice in education our "knowledge base" and our morals and politics should be integrated into an ideology of hope and sincere cooperation for a better future for children through education reform.

Overview

A characterization of a critical evaluator will be advanced shortly. The role divisions of academic versus service orientations existing in evaluation today are described. The ideology of a critical theory of education, and critical social psychology will then be reviewed in order to consider augmenting traditional positivist perspectives of evaluation. Afterwards I give a brief summary of evaluation in general. Three perspectives of evaluation and their purposes are explained, in order to illuminate the more traditional positivist approaches in prevalent current evaluation literature and to describe a spectrum of responsibility, purpose, and definition within the discipline of evaluation. The three perspectives on the spectrum are those of accountability, knowledge, and development.

Next, the limitations of contemporary and critical evaluation and how these approaches may implicitly serve a particular set of social and political values is forwarded. Integration of critical evaluation into a changing society, Fetterman's silent scientific revolution, the ideas of practicing critical evaluation, the neutrality of schools, and change in American schools are also presented. Subsequently this article conceptualizes one important process that an evaluator must experience in the context of Freirean pedagogy, so that a critical evaluator can begin the special role of critical evaluation in educational reform. The implications of critical social thought for evaluation in educational reform are then proposed. Finally, the second part of this article begins by describing the interdisciplinary methods and procedures of the "how" of critical evaluation, by introducing the integration of American public school law as enhanced by collaborative consultation and the adversary-advocate oriented evaluation model.

The Critical Evaluator

Ernest R. House was the first prominent evaluation theorist to advocate valuing based on principles of social justice (Patton, 1997). He has consistently voiced concerns for democratizing decision making in that context, he has analyzed the ways in which evaluation inevitably becomes a political tool in that it affects "who gets what." As mentioned earlier, education itself, as well as educational evaluation can enhance fair and just distribution of benefits or it can distort such distributions and contribute to inequality. In considering judgements on programs, the social justice evaluator, the critical evaluator, is guided by such principles and values as equality, fairness, and concern for the common welfare (Sirotnik, 1990).

Kenneth Sirotnik and Jeannie Oakes collaborated in this same endeavor by considering the epistemological connections between critical theory and evaluation. To be specific, they stated that if one accepts the proposition that inquiry is never value free and accepts social justice as the ethical starting and ending points for moral argument, then the accumulated body of work done by Freire (1973), Habermas (1971), and others points the way toward a useful epistemological synthesis, one that they called critical inquiry, that is evaluative by its very nature (Sirotnik & Oakes, 1990). By no means is critical evaluation a new idea. Regardless, the argument for fighting for social justice with critical evaluation of education reform is not a trivial one, but it is an argument that I have extended with much interdisciplinary literature and paradigmatic considerations.

Michael Quinn Patton (1997) wrote that social justice and other similar principles change the role of the evaluator from the traditional judge of merit or worth to a social change agent. Many evaluators surveyed by Cousins, Donahue, and Bloom (1996) were hostile to or at least ambivalent about whether evaluation, particularly a type of critical evaluation, can or should help bring about social justice. Certainly, evaluators undertaking such an approach need to be comfortable with and committed to it, and such an activist agenda must be explicitly recognized, negotiated with, and formally approved by primary intended users. From Michael Quinn Patton's utilization focused perspective, using evaluation to mobilize for social action and support social justice are options on the menu of evaluation process uses (1997).

In this article, part of the argument is that wherever one places oneself on the spectrum of evaluation responsibility, purpose, and definition; the evaluator can earnestly acknowledge the powerful critical role that he or she may interpret in placing judgement or giving merit to one of the most profound social activities in our lives, that of educating our students and our children. This role as a critical evaluator can be found anywhere on the spectrum. As typically happens with most spectrums the outlier situation is pretty rare. A critical evaluator can produce empirically traditional research designs in combination with critical social ideology, as long as one maintains a critical stance towards methods, practice, and policy that addresses the more difficult questions about the institutional problems in educational programs, those of democracy, power, and inequality. Patton (1994) also advocated the use of "mind shifts back-and forth between paradigms within one evaluation setting."

Most of the time, in most environments represented on the spectrum, "scientific" positivist traditions about knowledge and postmodern critical social constructions of knowledge are almost bound together, and evaluators must therefore always be prepared to confront them both (Young, 1990). In *Ethics, Politics, and International Social Science Research*, Hamnett, Kumar, Porter, and Singh (1984) compared and described the aforementioned theoretical presuppositions such as that of positivist constructions of knowledge and that of critical theories of knowledge. A significant point here is that a critical evaluator can utilize the necessary tools and methods within shifting research paradigms and changing concepts of knowledge construction, in order to augment practices and policies which are continuously participating in a discourse that values a moral imperative and a political commitment to social justice in the evaluation of education reform. This understanding of a moral imperative and a political commitment in educational evaluation is crucial in establishing explicitly the ideology of a critical evaluator and in making one's analytical biases clear.

The following paragraph provides a synopsis of Sirotnik's and Oakes' review of the three faces of inquiry and analysis (1990). Most educational researchers and evaluators have been schooled in the tradition of the scientific method and the hypothetico-deductive paradigm borrowed, presumably, from the physical sciences. But there are at least two other separate and general orientations for systematic inquiry having strong philosophical roots and demonstrable utility for the social sciences. The more familiar is the whole class of naturalistic methodologies. The second major departure from the empirical analytical tradition is less well known and much more separable, namely, the critique of knowledge. Its roots are also in the hermeneutical tradition. But as a philosophy of inquiry, it represents a significant extension of interpretive inquiry. Inquiry and analysis does not happen in a normative vacuum, as they so eloquently stated.

Sirotnik and Oakes (1990) also suspected that "an epistemological trap can be created through assuming necessary and sufficient connections between method and the political content of cognitive interests. Conducting empirical analytic inquiry, for example, does not necessarily imply a hidden agenda of domination. On the other hand, a hidden agenda of domination cannot in principle survive an inquiry based on critical theory" (p.45). I agree with these authors that this, indeed, points the way out of the trap, a truly practical unification of the three faces of inquiry requires the self correcting epistemological stance that is made to order in a critical perspective. At the same time evaluation must consider what kind of orientations are created in practice when these epistemological and empirical stances are postured.

Academic Versus Service Organizations

One of the most basic role divisions in the profession today is between academic and service oriented evaluators, a division identified by Shadish and Epstein (1987) when they surveyed a stratified sample of the members of the Evaluation Network and the Evaluation Research Society, the two organizations now merged as the American Evaluation Association. The authors inquired about a variety of issues related to evaluators' values and practices. They found that responses clustered around two contrasting views of evaluation. Academic evaluators tend to be at universities and emphasize the research purposes of evaluation, traditional standards of methodological rigor, summative outcome studies, and contributions to social science theory (Patton, 1997). Service evaluators tend to be independent consultants or internal evaluators and emphasize serving the stakeholders' needs, program improvement, qualitative methods, and assisting with program decisions (Patton, 1997).

According to Shadish and Epstein, " The general discrepancy between service oriented and academically oriented evaluators seems warranted on both theoretical and empirical grounds" (1987, p.560). The profession of evaluation remains very much split along, these lines, but with new twists and perhaps, deeper antagonisms (Patton, 1997). Patton goes on to explain how the "schism" erupted openly, and perhaps deepened, in the early 1990's, when morality entered into the evaluation arena much more explicitly, and the American Evaluation Association elected successive presidents who represented two quite divergent perspectives.

Yvonna Lincoln (1991), in her 1990 presidential address, advocated what Patton would call an activist role for evaluators, one that goes beyond just being competent applied researchers who employ traditional scientific methods to study programs, the academic perspective. She closed her speech by asserting that "my message is a moral one." The following year, the American Evaluation Association president was Lee Sechrest, who by his own definition represented the traditional, academic view of evaluation. He objected to Lincoln's metaphorical call for a new generation of evaluators. "I ask myself," Sechrest (1992) mused, "Where in our makeup are the origins of this new creation so unlike us.... I sense a very real and large generational gap" (p.2).

From this contemporary discourse in what the role divisions personify in evaluation, one can tell that critical evaluators may be characterized as divergent or even marginal in their theoretical and empirical presuppositions. Here lies the embedded professional challenge of remaining open to pluralistic and cosmopolitan approaches which adapt evaluation practice to new situations, mainly the situation of public education institutions which are failing a growing disproportionate amount of disadvantaged children thereby reproducing social and symbolic inequalities. Ultimately, there is no one way to conduct an evaluation. This insight is crucial. The design of a particular evaluation depends on the people involved and their situation.

Ideologies of Critical Theory, and Critical Social Psychology

Traditionally social science and social psychology we are told, is a vocation of scientific method, a devotion to truth that should not be compromised by the researcher's idiosyncracies or other external forces and should not be unduly affected by the social context in which the researcher operates (Hamnett et al., 1984). In the realm of the natural sciences, statements often appear to be reaffirming this stance. For instance, in practice there is very little to distinguish Soviet and U.S. nuclear physics. Changes in theoretical presuppositions in one country are rapidly translated to others.

Social science and social psychology, however, do not have the canons of theoretical perspective, verification, or even of data collection found in natural science (Hamnett et al., 1984). Hamnett and his co-authors state that theoretically, the sociology of knowledge has demonstrated how science (including the concepts, methods, and procedures embodied) presupposes historically relative values, interests, and ideologies. The taken for granted notion of the methodological neutrality of scientific method has been undermined by theorists of many persuasions including that of critical theorists and critical social psychologists (Wexler, 1983). I agree with Wexler when he writes that conventional wisdom and common sense concedes that science is influenced by human values and the political contexts of its expedition. This is why the evaluator of education reform cannot posture a neutral, purely objective point of view on the object of his research, especially in the reforming of such a social contract as education.

The writings of critical theory developed from the Institute for Social Research in Frankfurt. The critical theorists are concerned with the role of values and ideology as "part of the conceptual framework which defines what it is to have, i.e., scientific knowledge about some phenomenon" (Sabia & Wallulis, 1983). Such a focus raises important questions concerning social science research, ethics, and inevitably the practice of evaluation in education. Critical theorists state that it would be incorrect to claim that positivist doctrine is responsible for the unreflexive state of the research ethics and politics debate in social science; the social, historical, economic, and political context of research is of overwhelming importance (Sabia & Wallulis, 1983).

How one views the role of social research, its relations to political practice, and how one assesses responsibilities, relationships, and appropriate conduct should be explicitly negotiated up front with potential clients in terms of one's underlying assumptions and ideological presuppositions. Moreover, critical research methodology is distinctive from other approaches in that it traces the origin of our concept of validity back to everyday human interaction. This is true, at least, for the specific brand of critical methodology I advocate, which draws heavily from Habermas's work on validity (Habermas, 1981, 1987). The later discourse of this critical evaluation perspective, which can be embedded in a positivist scientific method, does not assume the posture of rejection or exclusion, but rather will serve to provide an additive component to constructing knowledge and representing it with critical and conscious eyes.

I repeat what Lewis Carroll's Alice would have said, "things are not what they seem." There is a difference between listening to the goals of reformers, and listening attentively to the underlying assumptions forwarded by education reform efforts, and consequently holding the reformers responsible for living up to their social ideals and their program mission statements, mainly those mission statements that become framed cultural symbols of what a program or a school represents. These framed paper certificates, these mission statements, are usually strategically placed in the front office of every public school and meticulously published in brochures summarizing the goals and objectives that school districts represent to welcome potential inhabitants of the communities they serve. If we can understand the central role played by validity claims in normal human communication (symbolic or otherwise), we will then be able to formulate the special requirements that a critical evaluator conducting formal inquiries into social processes must employ to produce a trustworthy account. In critical evaluation, the validity claims made by the evaluator do not differ in nature from validity claims made by all people in everyday contexts.

Critical social psychology draws from the critical theory of the Frankfurt School and the theoretical traditions of Marxism (Wexler, 1983). Philip Wexler (1983) augmented and amplified what he perceived as developing tendencies in social relations and in social psychological processes. Like Philip Wexler's expression of a need for augmentation, I am asking those who study, practice, and use evaluation in education to broaden and amplify their view of the applications and functions of evaluation with an eye to the future. The evaluator could be responsible for reaching beyond mainstream philosophy and practice in evaluation because the transforming of education and the reforming of such a significant social activity requires an exceptionally conscious human being. Like critical social psychology, a critical evaluator requires a theory which can comprehend and facilitate social change movements.

Next I shall give an overview of evaluation in general, its development, and then review three perspectives of evaluation and their purposes, in order to illuminate the more traditional positivist perspectives in prevalent current evaluation literature. These three perspectives again are those of accountability, knowledge, and development. By looking at these three perspectives and their positions along a spectrum, I argue that the evaluator must go beyond those delineated perspectives in mainstream evaluation theory, policy and practice, in order to take a more critical posture toward both education and the very process of thinking about education.

Evaluation

Evaluation as an academic discipline, a profession, and a government function has only developed in the past four decades in the United States and in several other industrially developed nations. In many nations, however, evaluation is in its infancy as a standardized pursuit; and certainly on a global scale, evaluation is only beginning to enter the scene (Chelimskv & Shadish, 1997). There is comfort in knowing, as previously mentioned, that current literature in evaluation confirms that evaluation as a social activity is becoming increasingly self- conscious about its own identity and purpose in the larger social order and is beginning to systematically evaluate its own methodology, utilization, values, and politics (Cronbach & Associates, 1980; Guba & Lincoln, 1989; Rossi & Freeman, 1993; Patton, 1994; Scriven & Kramer, 1995; Chelimsky & Shadish, 1997; House, 1993; Scriven, 1991). I would agree with Chelimsky and Shadish (1997) when they propose that evaluators, in whatever field of evaluation they may be, are likely to find themselves, at least sometimes, at odds with the political actors, systems, and processes in their own backyards, that rally against a free flow of information and collaborative action which endangers the status quo.

Between 1965 and 1990 the methodology, philosophy, and politics of evaluation changed substantially, partly in response to the structural transformations in an advanced capitalistic society (House, 1993; Scriven 1991). The strongest stimulus to the development of evaluation was Lyndon Johnson's Great Society legislation, which, though not capable of changing U.S. society as a whole, certainly transformed educational and social research. At Senator Robert Kennedy's insistence, the Elementary and Secondary Act in 1965 mandated evaluation of programs for disadvantaged students, and this spread to all social programs and beyond (McLaughlin, 1975). House (1993) reviewed clearly in lay terms how evaluation moved from monolithic to pluralist conceptions, to multiple methods, multiple measures, multiple criteria, multiple perspectives, multiple audiences, and even multiple interests.

Methodologically, evaluation moved from a primary emphasis on quantitative methods, in which the standardized achievement test employed in a randomized experimental control group design was most highly regarded, to a more permissive atmosphere in which qualitative research methods were acceptable. Mixed data collection methods are advocated now in a spirit of methodological ecumenism (House, 1993). The following three perspectives describe more thoroughly the way that evaluation is characterized in contemporary evaluation circles.

Examples of Purpose and Perspectives in Evaluation (Chelimsky & Shadish, 1997)

Below find a review of the definitions and characterizations that Chelimsky and Shadish write about in *Evaluation for the 21st Century*. They offer an inexhaustible listing of possible purposes for evaluation. These purposes include the following: (a) to measure and account for the results of public policy, and programs, (b) to determine the efficiency of programs and their component processes, (c) to gain explanatory insight into social and other public problems, (d) to understand how organizations learn, (e) to strengthen institutions and improve managerial performance, (f) to increase agency responsiveness to the public, (g) to reform governments through the free flow of evaluative information, and (h) to expand results or efficiency measurement from that of local or national interventions to that of global interventions such as reducing poverty and hunger or reversing patterns of environmental degradation. All of these purposes are, of course, worthwhile and legitimate reasons for conducting evaluations, but they differ with regard to the questions they address and the kinds of methods needed to answer these questions.

Chelimsky and Shadish propose that these different purposes, along with the questions they seek to answer, seem to fall naturally into three general perspectives:

- evaluation for accountability (e.g., the measurement of results or efficiency);
- evaluation for knowledge (e.g., the acquisition of a more profound understanding in some specific area or field); and
- evaluation for development (e.g., the provision of evaluative help to help strengthen institutions).

The methods of these three perspective are not mutually exclusive. Though they do represent notable differences on a variety of dimensions. Each may be needed at particular times or policy points and not others (e.g., evaluation for knowledge may need to precede accountability). Chelimsky and Shadish (1997) write that they appear to have considerable explanatory power with regard to the current tension in the evaluation field. (See Table 1 for further details.) Table 1, an adapted chart from Chelimsky and Shadish's book (1997, p.21), shows the following three different perspectives and their respective positions along five dimensions.

Table 1
Three perspectives and their positions along five dimensions, adapted from Chelimsky and Shadish (1997, p.21)

DIMENSIONS	ACCOUNTABILITY PERSPECTIVE	KNOWLEDGE PERSPECTIVE	DEVELOPMENTAL PERSPECTIVE
PURPOSE	to measure results or value for funds expended: to determine costs, to assess efficiency	to generate insights about public problems, policies, programs, & processes, to develop new methods and to critique old ones	to strengthen institutions to build agency or organizational capability in some evaluative area
TYPICAL USES	policy use, debate and negotiation, agency reform, public use	enlightenment use, policy, research and replication, education, knowledge base construction	institutional or agency use as part of the evaluative process, public and policy use
EVALUATOR ROLE	distant	distant or close depending on evaluation design and method	close, the evaluator is a "critical friend" or may be part of a team
ADVOCACY	unacceptable	unacceptable, but now being debated	often inevitable, but correctable through independent outside review
POSITION UNDER POLICY DEBATE	can be strong (depending on leadership)	can be strong (if consolidated and dissemination channels exist)	uncertain (based on independence and control)

The Accountability Perspective

From the standpoints of auditors, government sponsors of evaluation studies, donors to international organizations, and many others, evaluation is done to establish accountability. This involves the provision of information to decision makers, whether they are in the public or private sector. Specific cause and effect questions about the results in an accountability perspective might be: What happened to poverty levels among the very poor as a result of development assistance provided" Did an educational intervention or program produce more "effective" learning for all learners? Has teacher training increased student achievement?

Sometimes, questions about the results from an accountability perspective may involve merely documentation of whether or not anything has changed after something new has been tried. Normally, however, the ability to say that something is in fact a "result" hinges on the ability to establish that it came about because of something else. Many methods are used to answer these kinds of accountability questions including: randomized designs, quasi-experimental designs, mixed multi-level designs, mixed qualitative/quantitative designs, case studies, process studies, and research synthesis designs.

The Knowledge Perspective

In the view of many researchers working independently in universities and other evaluators in scientific institutions, evaluation is done to generate understanding and explanation. Chelimsky and Shadish (1997) stated that the specific questions may not be especially important to analyze here, given that it is the evaluator who decides what will be asked and answered, and the topic generally follows from the researcher's prior work. They explained that the evaluations associated with individual academic researchers, or those of research teams, will be more likely to continue in depth cumulative inquiry into particular areas or sectors of research than to be concerned with applying systematic research methods to a variety of sectors, as with accountability and developmental evaluations.

The larger purpose of the knowledge perspective is to increase understanding about the factors underlying public problems, about the "fit" between these factors and the policy or program solutions proposed, and about the theory and logic (or lack thereof) that lie behind an implemented intervention. "These evaluations may employ any of the methods discussed above, separately or in conjunction with each other, but the purpose of knowledge gain leads logically to the use of the strongest designs as well as the greatest clarity possible in explication and documentation of methods to facilitate replication or later use in research synthesis and policy formulation" (1997, p.14).

The Developmental Perspective

For government reformers, public managers, and others, evaluation is done to improve institutional performance. It serves as a flexible tool that works: (a) to improve the design of projects, (b) to measure and recommend changes in organization activities, (c) to develop the indicators and performance targets needed to improve institutional effectiveness and responsiveness, (d) to monitor, in an ongoing way, how projects are being implemented across a number of different sites, and/or (e) to find out how beneficiaries feel about an agency and its programs. To some accountability or knowledge perspective evaluators, developmental evaluators may seem more like evaluation "consultants" than evaluators, but those who do developmental work are convinced that building evaluation capability is as important an evaluation function as evaluation itself and that indeed, in some cases, evaluation cannot be done without capacity building.

Specific questions asked of evaluators in a developmental perspective might include the following: What is the best research evidence with respect to formulating a new program or modifying an old one? How can projects be structured so that they produce evidence on the value of the intervention being tested? What is the most appropriate agenda for the agency? Both process and outcome designs may be used in a developmental perspective, depending on the evaluation question posed. In addition to the methods mentioned earlier, the formative methods used in the developmental perspective include the following: monitoring, empowerment evaluation, cluster evaluation, performance measurement, and research synthesis of both qualitative and quantitative methods. A developmental evaluator becomes part of the design team helping to shape what's happening both processes and outcomes, in an evolving, rapidly changing environment of constant interaction, feedback, and change. Using mixed methods and multiple criteria in this perspective are productions of some of the many current trends in the practice of evaluation.

Demonstrating a Particular Set of Social and Political Values

Although evaluation has developed as a discipline, a profession, and as a government function in the past four decades by building on its scientific positivist traditions and by systematically evaluating its own existence in the larger social order, this particle emphasizes continual growth and augmentation of its practices, policies, and responsibilities through a "conscientization" of evaluation's socio-political reality. Over the years evaluation has come to be seen as political. Michael Quinn Patton, at the National Evaluation Conference in Youngstown State University held in September 1998, summarized 12 recent trends in evaluation. One of them being the increasing political sophistication and acknowledgment of the role of values and morals in evaluation practice. There can be no doubt, that evaluation is influenced partly by political forces, and in turn, has political effects. Whose interests are served and how interests are represented in an evaluation are now very critical concerns in a society with increasing disparate value systems.

In the earlier days, it was assumed that the interests of all parties were properly reflected in the traditional outcome measures, but this assumption came to be questioned, and it was recognized that different groups might have different interests and might be differentially affected by the educational program and its evaluation (House, 1993). "Stakeholders" (those who had a stake in the program under review) became a common concept, and representing stakeholder views in the evaluation became an accepted practice.

The stakeholder concept is based on the prevailing pluralist-elitist-equilibrium theory of democracy, which disclaims any normative judgements and which holds that the current system of competing parties and pressure groups performs the democratic function of equalizing the diverse and shifting political demands (MacPherson, 1987). It is perceived that describing what others value is the stance best suited to the political context in which evaluators operate, because decision making depends on the values held by relevant policy makers and stakeholders. Presumably, these parties will use the findings to make informed decisions. Neither the government nor the evaluator is supposed to intervene to support any particular interests but rather only to provide information that is value-neutral and interest-neutral. The interests of various groups somehow dissolve into the values of decision makers and stakeholders.

However, one must note that today's professional evaluators sometimes become evaluators by default. We represent an eclectic and diverse combination of various professional, academic, and research areas. Shadish and Epstein (1987) found that 31% of the respondents in their survey described their primary professional identity as that of "evaluator" (p. 560). Others thought of themselves first as a psychologist, sociologist, economist, educator, and so on, with identity of evaluator secondary. When both Charles Murray (1983,1984) and Michele Fine (1983b, 1988) have been successful evaluators representing a particular set of social and political values and interests, one has to acknowledge the diverse socio-political reality in which evaluators actually find themselves in practice.

In two highly visible stakeholder evaluations funded by the federal government, those of Cities-in-Schools and Jesse Jackson's PUSH/Excel program, the evaluations worked against the interests of the program participants and the inner-city students which the programs were supposed to serve, thus calling into question the justice of these evaluations (House, 1988; Stake, 1986). The results of the PUSH/Excel evaluation were used not only to discredit the program but also to question Jesse Jackson's ability to manage large enterprises during ensuing presidential campaigns. In truth, the stakeholder model was never implemented (House, 1988; Stake, 1986). Charles Murray, the evaluator in both cases, substituted a technocratic model of evaluation and expressed his disdain for the stakeholder concept in his article *Stakeholders as Deck Chairs* (1983). Although the stakeholder approach seems firmly entrenched, there is disagreement about how to implement it. In reality, stakeholders do not have equal power to influence and utilize the evaluation, nor do they have equal protection from the evaluation.

These types of problem in evaluation led into a discussion of misuse of findings. The fact that so much standardized achievement testing is reported to the public in general and its interpretation left to the media or government officials makes misuse particularly salient (House, 1993). In fact, the professional standards for evaluation developed by a committee led by Stufflebeam, devoted considerable space to issues of misuse, but the context in which evaluation results are presented does not lend itself to the employment of such standards, even though the standards are widely accepted in the evaluation community itself. How misuse of findings and disparate interests can be curtailed is by no means clear. The professional standards for evaluation developed by the Joint Committee dramatically reflected the ways in which the practice of evaluation had matured in 1981. In 1994, revised standards were published following an extensive review spanning several years.

While some changes were made in the 30 individual standards, the overarching framework of four primary criteria (utility, feasibility, propriety, and accuracy) remain unchanged. However, the profession of evaluation has not yet developed to the point of reflecting a common core of practices and principals as demonstrated by the original professions, divinity, law, and medicine (House, 1993). We must pay attention to the fact that certification programs and higher education programs in evaluation and evaluation research are a very recent development in the discipline (Chalimsky & Shadish, 1997). For a deeper understanding of how the original professions compare with evaluation as a profession, refer to House's (1993) book, *Professional Evaluation* .

Limitations of Contemporary Evaluation and a Reflection on American Public School Law

There are limitations to contemporary and critical evaluation frameworks. The problem of addressing multiple values and interests and how they should be represented in an "equitable" evaluation can take one directly into the realm of social justice and the recognition of the assumptions, character, and consequences of conventional forms of educational evaluation and American public school law. The problem of evaluation representing a particular set of political and social values (i.e., a broadly conservative set) also raises some serious questions about evaluation in general. Although the socio-political reality of multiple stakeholders and evaluators who have legitimate values and sometimes conflicting interests is recognized, how these values and interests are legitimized will become one of the most important challenges for educational evaluation in the future, especially for critical evaluation of education reform. How to synthesize, resolve, and adjudicate all these multiple multiples in our increasing multicultural and amorphous society remains a formidable question, as indeed it does for the larger society.

One thing we do know is that the socio-political reality in evaluation of public programs, both in education and health, often works in favor of higher income groups and against equity despite the stated objectives (Birdsall & Hecht, 1995; Paul, 1991; Fine, 1983). When we look at the political structures and the broad organization of society, resource allocation and subsequent delivery of services and programs tend to be skewed in favor of those who have more "voice" (Fine, 1983; Fine & Weis, 1993). In many cases, powerful stakeholders or groups, which are able to effectively demonstrate their interest in receiving social services and "effective" or "successful" social programs, manage to get the lion's share of the resources and the funds. It is no secret that the United States of America is one of the last Western industrialized nations to base their educational financing system on that taxation of largely differentiated property values. This financial arrangement alone should illuminate some of the deeper issues at stake in the evaluation of public education environments.

American public school law and its case history has demonstrated time and again that there are very few instances where citizens have been able to prove that state school finance systems result in revenue disparities which violate the Equal Protection Clause of the Fourteenth Amendment. In 1973, in the case of San Antonio Independent School District v. Rodriguez, Mr. Justice Powell, delivered the opinion of the Supreme Court. He said, "to the extent that the Texas system of school financing results in unequal expenditures between children who happen to reside in different districts, we cannot say that such disparities are the product of a system that is so irrational as to be invidiously discriminatory...."

If disparate allocation of governmental benefits can be justified on the basis of reasonable classification or the interests involved are not fundamental, then statutes will be regarded as constitutional (Alexander & Alexander, 1992). The court in the Rodriguez case basically ruled that a state legislature can heap benefits on some wealthy school districts and deprive others of fiscal resources and not offend the federal Equal Protection Clause. Thus representing the educational interests of disenfranchised stakeholders, even within the American public school law domain, can be confounded with many inherently unequal and disparate value systems.

In other instances, our social service institutions, such as education and health, are able to shape the systems to serve their own personal and professional goals at the expense of equitable delivery (Paul, 1991). Problems created by the limited voice of politically weak or disenfranchised stakeholders are exacerbated in educational evaluation, when combined with direct provision of services in virtual public monopolies of the "best teachers," the allocation of "best practices" in education, and the provision of high quality curriculum and professional development training which are centralized in higher socio-economic communities. Ultimately, citizens have limited capacity to improve the public education they are provided through participating, informing, and making recommendations. This is especially true of lower socio-economic community stakeholders which have traditionally been limited in their capacity to have their "voice" heard without legal representation (Fine, 1993; Oakes & Guiton, 1995).

Historically, when interests have been ignored and educational procedures have been violated, lower socio-economic communities, minorities, exceptional populations, and limited English proficient citizenry have had to turn to the legal system for any kind of adjudication (Paul, 1991; Haring, McCormick, & Haring, 1990; Oakes & Guiton, 1995). Similarly, in terms of fighting for social justice in education, evaluation of education reform efforts could benefit from addressing some of the principals in American public school law. This idea will be further developed in Part II below. However, for the time being, contemporary evaluation which was invented to solve social problems, can be afflicted with many of the problems it was meant to solve.

Another limitation of critical evaluation in education reform pertains to its inherent questioning of the institutional character of education. By producing educational criticism and value judgements of institutional programs and personnel, in conjunction with the ideologies of critical theory of education, critical social psychology, and Freirean pedagogy, critical evaluators risk certain professional isolation from the mainstream. The socio-political reality in which one can survive as an evaluation professional of education reform becomes integrated into a world with those individuals that agree with your views, particularly those who agree with your views on social justice and in general the democratic purposes of public schooling. As critical evaluators conduct evaluations to address the elimination of inequality and the improvement of those who are least well off, they will come into conflict and threaten established authority. Any method of evaluation that claims to be nonobjective and value-laden will be marginalized. Society expects evaluation to be based on scientific authority. However, I expect the notion of what is scientific to be substantially redefined. The concepts of objectivity, scientific methodology, and validity will be recast to accommodate different evaluation approaches (House, 1993).

Integration of Critical Evaluation into a Changing Society

Evaluation continues to become ever more methodologically diverse. Evaluation in general draws from the theoretical foundations of many fields and is multi-disciplinary and multi-faceted in nature (Chelimsky & Shadish, 1997). It is by now well established that the full array of social science methods belongs in the evaluator's methodological tool kit, including tools from psychology, statistics, education, sociology, political science, anthropology, and economics (Cronbach & Associates, 1980). When the critical logical and analysis tools given to us by critical theorists and social psychologists are included into an evaluation design, the role that evaluators play in judging the worth of educational reform efforts is elaborated. Chelimsky and Shadish (1997) supported the notion that it is often uncomfortable to stir oneself from familiar cultural, ideological, topical, conceptual, and methodological niches.

However uncomfortable or reactionary one may feel to the content of this article, there is a message: it is that evaluations of educational reform efforts in the next century can and probably will be far more powerful and influential than they are today. This is because of the ever increasing complexity of social, economic, technological, political, and cultural tensions which are questioning the very integrity and purpose for public education as a whole (Giroux & Aronowitz, 1991). The growing populations with disparate value systems and socio-economic levels and the increasing minority populations in this country will demand to participate more legitimately in the reformation of their own education. Consequently evaluation will have to redefine its identity, its purpose and practices.

Lee Cronbach, in 1980, advanced the position that the theory of evaluation has to be as much a theory of political interaction as a theory of how to determine facts or how knowledge is constructed (Cronbach et al., 1980). Even so, 18 years later, we still do not seem to understand political processes very well, especially their dynamic nature. This gap in understanding and consciousness is especially true for evaluators in the field of education where we are determining "facts" and constructing knowledge about educational programs designed to improve teaching and learning in the public school domain. We can begin bridging this gap in consciousness to understand the political nature of evaluation by looking at our own ideologies as evaluators. Critical thought, indeed, criticism, is essential to enable us to act in ethically and politically just, to say nothing of intellectually honest, ways. Critical thought entails questioning, reflection and thoughtful interaction with the information and body of knowledge at hand. Education then becomes an active and constructive process of continual critical growth (Dewey, 1944).

Fundamentally, I am recommending here that evaluators, as leaders of educational reform efforts, become more critical and vigilant about the questions they are contracted to answer and about the more profound functions of education programs and practices under the rubric of a critical theory of education (Giroux, 1983b; Young, 1990; Apple & Beane, 1995; Apple 1996, Apple & Carlson, in press). In addition, these same evaluators could integrate the logic of traditional psychology with the logic of critical social psychology to begin the rethinking of education as a social project and a social process. The purpose of this rethinking is to expand on positivist traditions of considering an incremental perspective on methodological and research design issues in evaluation, into a more open critical ideology of practice and policies (Fetterman, 1988; Maruyama & Deno, 1992). These schools of thought, approaches, and particular issues should not be eliminated. We should consider these issues together with the notion that evaluation of education places us in a particularly sensitive arena within the confounds of social and human science (Fetterman & Pitman, 1986; Fetterman, 1988).

Silent Scientific Revolution

Fetterman (1988) argued that there is a silent scientific revolution in evaluation and that educational evaluation is experiencing a change in direction. A critical component of this change is a shift in the paradigms underlying the method and aim of research (Lincoln, 1986). David Fetterman further suggests that a marked shift is taking place in the professional allegiance of evaluators. This shift in allegiance, he says, is not a simple linear development. As summarized in Fetterman's book (1988), this shift goes beyond perceiving evaluation as a set of chronological transformations that travel from traditional positivist approaches to dominant qualitative forms of evaluation, including ethnography, naturalistic inquiry, generic pragmatic (sociological) inquiry, connoisseurship/criticism, and phenomenography. Rather he illustrates some significant moments of metamorphosis, revealing the process of shifting allegiance to a circular and interactive paradigmatic perspective.

Similarly, I call on evaluators to lift the blinders of methodological habit, to increase the ideological options and backgrounds available to them, to go beyond any single discipline, and to build on tradition by engaging the wisdom of critical social thought. This article is simply describing a possible interplay between the sciences and between the contemporary perspectives in evaluation.

Whether using the perspectives of accountability, knowledge or development, or any combination thereof, additional questions could be examined as the evaluation/research design is imposed on the school culture and setting (Maruyama & Deno, 1992). Critical evaluators of education reform could also listen to emerging questions that are integral to the improvement and restructuring of social projects and social processes, by attending to their own consciousness and motivations (Young, 1990). Later I will review Paolo Freire's construction of *conscientization* (Hamnett, Kumar, Porter, & Singh, 1984, p.100) to describe this experience as necessary for critical evaluators of education reform.

School and university researchers/evaluators who are taking on the challenge of restructuring schools and school systems in urban areas are involved essentially in the transformation of existing bureaucracies, bureaucracies that have had the power to control what is taught and how schools are run (Kretovics & Nussel, 1994). Clearly American education is organized in a bureaucratic form. Kretovics and Nussel confirm that at any level, national, state, or local, the traditional pyramidal, hierarchical arrangement is in effect. Proposals for reform of public schools and their evaluations must consider how the bureaucratic functionaries might respond. Since bureaucracy is "an institutionalized method of organizing social conduct in the interest of administrative efficiency," the issue of response is a genuine concern (Kretovics & Nussel, 1994).

Practicing Critical Evaluation

In the public school domain, genuine concern is adequate but more critical thought and action are needed within the world of educational bureaucracy. One way of practicing critical thought and action for critical evaluation would be to negotiate these ideological and theoretical presuppositions up front with one's clients and then deliberately confront issues of institutionalized power, democracy, and inequality in the educational programs and reform efforts. One can do this by organizing specific research designs and relationships centered around the concept of listening to the multiple voices in education and its programs. Fine and Weis (1993) witnessed and wrote about the practices and consequences of silencing in public schools. I do not think that evaluators are far from becoming partners in these implicit practices. Battling the dynamics of power and privilege that nurture, sustain, and legitimate silencing in education is the first purposeful step that a critical evaluator can take to interpret his powerful role as a transformative agent for social change. Creating flexible, authentic, and reflexive relationships with the stakeholders and with the existing bureaucracies during the process of evaluation is a second step that the critical evaluator can take towards completing a critical evaluation (Schon, 1983).

If innovative and well meaning educational programs or educational reform efforts are developed to improve the education of all students in public schools, then the evaluator of these programs has a very special and conscious role in creating opportunities for authentic discourse about these difficult issues that go beyond the successes or failures (outcomes) of children within the structural and organizational components of educational practice. The role of the evaluator and the ability to communicate and address the challenging issues such as democracy, power, and inequality to clients in the field of education, especially in the future, will be essential to transforming social activity for social change. Michelle Fine and Lois Weis (1993) included the following quote in their book:

> It is a false dichotomy which suggests that academics and/or intellectuals can only speak to one another, that we cannot hope to speak with the masses. What is true is that we make choices, that we choose voices to hear and voices to silence. If I do not speak in a language that can be understood, then there is little chance for dialogue. We must be ever vigilant. It is important that we know who we are, who we are speaking to, who we most want to hear us, who we most long to move, motivate, and touch with our words (p.2).

The Neutrality of Schools (Social Darwinism Revisited)

Jeannie Oakes (1986) stated that in their general indictment of schools, the authors of the evaluation studies and reform reports do not attach particular importance to the fact that schools fail to serve all students equally well. Certain topics like institutionalized power, democracy, and inequality are not explicitly addressed because there is a "silenced" understanding of the status quo in educational practice. Consequently, the evaluators and reformers in the commentary made by Jeannie Oakes do not consider as targets of information or understanding the school content and processes that limit school achievement for poor and minority students. Schools, in general, are often seen as essentially neutral, and the reforms are presented as color-blind and affluence blind. Jeannie Oakes (1986, 1995) further argued that current reform efforts do not address the unequal quality of school facilities, programs, materials, counseling, expectations, and instruction. No interest is shown, for example, in the unequal distribution of competent teachers. Neither do they address school organizational changes likely to equalize access to high quality educational contexts such as desegregation, the elimination of tracking, and the reconceptualizing of vocational education programs.

Thus by extracting the logic of critical theory and critical social psychology, I would extend the "meta-evaluation" done by Oakes, in saying that the evaluators of these reform efforts are additionally hard pressed to face squarely the "silent" demons lurking behind the institutional practices in public education. Even though a common issue is made of increasing the skills and knowledge of teachers, the assumption is that teachers simply need to get better at what they've always done. Also there is an assumption that all the evaluator has to do is to evaluate how the teacher is teaching and whether the outcomes are effective learning. There is little or no mention of the need for teachers to be more knowledgeable about how poverty, racism, and limited expectations affect the educational treatment of poor and minority children (Levine, 1971; Coleman, 1981; Fine, 1983,1994). Indeed there is no direct mention and acknowledgment of these issues on any explicit level within the hierarchical structure and bureaucracy in education (Levine, 1971; Coleman, 1981; Fine, 1983,1994).

Subsequently, mainstream evaluation of these reform efforts in teaching practices and educational programs misses a crucial part of the picture about how schools are functioning for all children. If we as evaluators do not ask deliberate questions about institutionalized power, democracy, quality of instruction, and inequality within the public school domain, during the process of evaluation, then we become one more vehicle that perpetuates an already neutral state of mind about the world of education and its goals for society. While many faults are found with schools, unfairness is not one of them. In addition, the omission of these concerns and "silent" demons in evaluation and education reform efforts makes clear the prevailing conviction that schools, as they are now, are neutral places (Coleman, 1975; Oakes, 1986; Fine, 1994).

Change in American Schools

Although there is a perception that change needs to occur in virtually all American school districts, including those serving the wealthiest suburbs, the success of the reform movement will be measured ultimately by its impact on our largest most troubled public school environments. For it is in our largest cities and our most rural districts that the job of the schools is most difficult, given the often overwhelming social and economic circumstances of students living in desperately impoverished neighborhoods (Oakes & Sirotnik, 1986). These are the neighborhoods most in need of transformed schools, and it is in these neighborhood schools that the evaluator can choose to undertake his exceptional role of being a vehicle for change and transformation.

Jonathan Kozol in *Savage Inequalities* (1991) took readers inside schools in poor neighborhoods and forced them to see the places impoverished children are compelled to go. Kozol (1991) commented on more than the physical, economic, and social inequalities among different types of school, those with affluent children, and those with children from poor homes. He addressed the very "ethos" of a school as maintained by the social-class position of the students. Theodore Sizer in *Horace's Compromise* (1984) also characterized this difference between schools quite modestly:

> Among schools there was one important difference, which followed from a single variable only: the social class of the student body. If the school principally served poor adolescents, its character, if not its structure, varied from sister schools for the more affluent. It got so I could say with some justification to school principals, tell me about the income of your student's families and I'll describe to you your school. (p.6)

Critical educators such as Michael W. Apple, Henry A. Giroux, Paolo Freire, Jeannie Oakes, Gloria Ladson-Billings, and Maxine Greene would probably agree that evaluation and research in impoverished neighborhood schools presents the critical evaluator with an exceptional challenge in social responsibility. Hence, these impoverished neighborhoods, where educational reform proponents advocate change, improvement and restructuring of schools, could be the environments that create wonderful opportunities for evaluators to maintain a critical stance toward theory, research, practice, and social policy.

Freirean Pedagogy

The statement "All men are created equal" is one that resounds throughout American history. The words are found in the Declaration of Independence and Lincoln's Gettysburg Address; they are also paraphrased and applied in numerous settings. For educators and educational evaluators, it has meant that American schools are charged with offering every child equality of educational opportunity. This concept of equality of educational opportunity is one that has been implicit in most educational practices throughout the period of public education in the nineteenth and twentieth centuries (Coleman, 1981). However, no white suburb in America would long tolerate the low academic achievement taken for granted in the urban, or rural public schools attended largely by African- Americans, Hispanics, and poor children.

In big cities all over the United States, minority students by the tens of thousands leave school each year, some as dropouts, some as graduates, utterly unprepared to participate in and contribute to a democratic society (Oakes & Sirotnik, 1986). They lack the skills that will allow them to obtain gainful employment, and they are devoid of the preparation that will lead to success in further education. Paolo Freire would characterize this lack of skills and preparation as the "inability to act upon and transform one's world" (Hamnett et al., 1984). Consequently he would say that the democratic society failed to move this person toward the ever-new possibilities of a fuller and richer life individually and collectively through the auspices of public education (Hamnett et al., 1984).

Paolo Freire is most often recognized for his literary and practical works as an educator. His study and conduct in this field have produced radically new philosophical and political insights. His basic assumption is that people are seen to be always in the process of developing. He says that the characteristic of the human species is its repeatedly demonstrated capacity for transcending what is merely given, what is purely determined (Hamnett et al., 1984). From Freire's point of view, education, or any form of activity directed at social change, can never be neutral; it can only be used to dominate or liberate people. Although this dichotomy is limited in my opinion, these extremes serve their purpose in explaining unique ideological commitments to social change, especially as social science researchers and evaluators in education. I proposed here that evaluation of public educational programs, as a form of activity directed at social change, should follow Freire's recommendation for *conscientization*:

> Conscientization refers to the processes in which men, not as recipients, but as knowing subjects, achieve a deepening awareness both of the socio-political reality which shapes their lives and their capacity to transform that reality (Freire, 1970b).

This notion conveys the realization that nobody can help or assist others without their participation; otherwise the helper is led only to treat people as objects vulnerable to control and manipulation from outside (Freire, 1973). Here we can reflect upon what such a perspective would require in evaluation of public educational programs. Conscientization is at least one experience that critical evaluators should pass through in order to become educational leaders and change agents for educational reform.

Implications for Evaluators of Education Reform

Undoubtedly the purposes, methods and functions of evaluation would change if one was to adhere to the philosophical and ideological underpinnings of critical theory, critical social psychology, and Freirean pedagogy. The question remains: would a critical evaluator actually go beyond traditional methodological concerns to design his policy and practice to deliberately address difficult and possibly uncomfortable issues such as institutionalized power, democracy, and inequality in education? Courage, persistence and conviction are presented here as three crucial elements that will consistently be needed for critical evaluation of educational reform. In addition to these three elements a critical evaluator could benefit from continual reflection about one's own changing beliefs and landmark experiences.

The need for courage, persistence, and conviction seems fairly obvious but somehow we do not seem to talk about these character traits explicitly. Speaking out, in situations that may include numerous political and bureaucratic agendas, all with different viewpoints and axes to grind, and also insisting on the right to independence in speaking out, takes a strong stomach. Even in the political and cultural environments occurring toward the middle of the spectrum, the normal skepticism of the evaluator is unwelcome amongst the pervasive enthusiasm for one program or another. But as we move down the spectrum toward differing ideologies, doubting the conventional wisdom becomes such an offensive tactic as to deconstruct credibility and solid reputations.

It also takes fortitude or conviction and strong resistance not to succumb to political or bureaucratic blinders of one sort or another. In my experience with the higher echelons of public education evaluation, both as a teacher and as a district based advisor of educational practice, these blinders lure evaluators into wanting to become political and "institutional" players on the national scene. There is an insidious temptation to avoid ideological and philosophical battles to the promise of glorious career rewards as compensation for obedience. This possible temptation is one of the reasons why there needs to be extensive research started in discovering the implicit and explicit social and professional ethics of different types of evaluators, especially evaluators in education reform. It takes persistence and courage to refuse sponsors the answers they want to hear, and it takes conviction and certainly conscience to ask deeper more resounding questions. Goethe said, "Possessions lost, nothing lost. Principles lost, something lost. Courage lost, everything lost" (quoted in "Visions of Public Service," 1986, p.12).

A Beginning to the Methods of Critical Evaluation

Part II: The "How" of Critical Evaluation

National policymakers, educational leaders, "public intellectuals", and children in disadvantaged situations can benefit from critical evaluation, but not in the same ways and not with the same evaluator roles. Neither more nor less activism, in my judgement, is morally superior. Various degrees of activism involve different ways to practice as an evaluator, often in different arenas. Indeed, how activist to be, involves consideration of an evaluation's purpose, decisions about intended users and uses of evaluation, and the evaluator's own values and commitments, all of which need to be made explicit. The challenge will be to create appreciation and space for such diversity among both those within and outside the profession who have a single and narrow view of evaluation and its practice. The debate will and should, go on, for that is how we discover implications and ramifications of diverse approaches, but I hope and foresee no desire to turn back the clock to a single dominant perspective.

By now, there should be no doubt as to the rationale for making a space for critical evaluation in the reform of public education. Because of the complexity of the task of reconceptualizing the evaluation process toward a process that contains an explicit normative social goal, that of social justice, and a process that is designed for purposes of fundamental change, the arguments in this section will only *begin* to delineate a preliminary path toward a methodology for critical evaluation. However, a more detailed and experienced methodology for critical evaluation would require further conceptual and empirical investigation and time. Essentially the utilization of American public school law, both state and federal statues, are combined with the adversary oriented evaluation model in order to propose briefly that these statutes can serve as merit criteria for determining the value and worth of educational programs. Critical evaluation will be augmented by commissioning the principles and rules of American public school law as additional references. Lastly, the conclusion elaborates on the roles and responsibilities of an evaluator in order to highlight the significance of our commitment and vision.

Adversary Oriented Evaluation (AOE)

Adversary Oriented Evaluation is a rubric encompassing a collection of divergent evaluation practices. In its broadest sense, the term refers to all evaluations in which there is planned opposition in the points of view of different evaluators or evaluation teams, and a planned effort to generate opposing points of view within an overall evaluation. In 1965, Guba suggested that educational evaluation might use aspects of the legal paradigm. I am suggesting not only to use certain aspects of the legal paradigm, but also to use the state and federal statutes as merit criteria for determining the worth and value of educational programs, especially those instructional programs that serve disadvantaged students.

Next, Worthen, Sanders, and Fitzpatrick (1997) presented a provocative rationale for such an approach. If trials and hearings were useful in judging truth claims concerning patents, products, crimes, civil disobedience, and if human testimony were judged acceptable for determining life or death, as in the judicial system, then might not legal proceedings and public education law be a useful metaphor for educational program evaluation? Might there be merit in educational evaluation "trials," in taking and cross-examining human testimony, and in using the concept of advocacy to ensure that evaluation fairly examined both sides of issues?

The first self-conscious effort to follow a particular adversary paradigm was made in the early 1970's by Owens. Designed to test the usefulness of a modified judicial model, the evaluation focused on a hypothetical school curriculum and included pretrial conferences, cases presented by the "defense" and "prosecution," a hearing officer, a "jury" panel of educators, charges and rebuttals, direct questioning and redirected questions, and summaries by the prosecution and defense (Worthen et al., 1997). The reports (Owens, 1973) were intriguing to the community of evaluators and led to further conceptual and empirical work on the adversary approach. For further explanation of the development, applications, strengths, and limitations of this kind of approach see Worthen, Sanders, and Fitzpatrick (1997).

Several approaches that qualify as adversary oriented do not employ hearing processes. Kourilsky and Baker (1976) described an adversary model in which two teams prepared, respectively, affirmative and negative appraisals of that which was evaluated (the preparation stage); met to present the views to one another, cross-examining and critiquing one another's contentions on pre-specified criteria (the confrontation stage); and engaged in open-ended discussions until reconciliation of views was attained and translated into written recommendations in a single report. Levine (1974) proposed that a resident adversary or critic might be assigned to the research project to challenge each bit of information collected, searching for other plausible explanations. The Stake and Gjerde (1974) strategy of having two evaluators prepare separate reports summing up opposing positions for and against the program is yet another variant of the adversarial approach that does not depend on a hearing format. These proposals are all consistent with what Worthen et al. (1997) also called "critical evaluation."

Donmoyer (n.d) proposed a "deliberative" approach to evaluation, which focused on assessing and balancing alternative conceptions of reality and the differing value positions underlying these conceptions. "Because deliberative evaluation is primarily concerned with fostering understanding of alternative conception of reality," the evaluator's role is "to foster interaction and facilitate communication among representatives of various stakeholder groups...." (p.9-10). Donmoyer saw different world-views as the cause of underlying disputes, which could be resolved by open presentation of alternative views in some type of forum.

Worthen et al. (1997) reviewed three general approaches to adversary evaluation: (1) adaptations of the legal paradigm and other "two-view" adversary hearings, (2) adaptations of quasi- legal and other adversary hearings where more than two opposing views are considered, and (3) use of debate and other forensic structures in adversary evaluation. The third type is particularly interesting for critical evaluation purposes of establishing merit criteria using the public education laws and codes that can serve as partial "anchors" or references for determining the quality of instructional and educational program delivery. The following is a practical representation of how the education laws and codes can be used as partial "anchors" or references.

For example, if the instructional effectiveness of programs such as bilingual education or special education was to be evaluated at a predominantly Hispanic low socio-economic elementary school in Texas, the critical evaluator could turn to the Texas Law School Bulletin (1996) for crucial information on the state's public education laws and codes that applied to the "Educational Programs" (Chapter 29, Subchapters A & B). A critical evaluation could include an investigation of the history of eligibility, assessment, enrollment, and placement into the bilingual and special education programs as defined in the Texas Law School Bulletin. Similar to the study completed by Jia Wang (1998), as mentioned previously in this article, the evaluation design would also include investigating the quality of instructional delivery, content coverage, content exposure, and content emphasis (opportunity to learn variables as described by Jia Wang, 1998).

In some instances, if the educational development of certain disadvantaged students, such as their language proficiency and academic achievement or failure were called into question, the evaluation team could review carefully the student's educational history by comparing it to the eligibility criteria, assessments, enrollment, and instructional placement education codes as set out by the Law Bulletin. These education codes could be the "anchors", the starting points or references to further the understanding of current and past campus and district based educational practices that involve high risk decision making. Education code 29.056, Enrollments of Students in Program is an example of this kind of "anchor" or reference:

> The agency shall establish standardized criteria for the identification, assessment, and classification of student of limited English proficiency eligible for entry into the

> program or exit from the program. The student's parent must approve a student's entry into the program, exit from the program, or placement in the program. The school district or parent may appeal the decision under Section 29.064 (p. 120).

Again, the laws and codes can be used as additional references for the evaluators to place classroom instruction, the school, the program, or the school district, in context to legal precedent and required administration. Because a public school is a governmental agency, its conduct is circumscribed by precedents of public administrative law supplemented by those legal and historical traditions surrounding educational organization that is state established, yet locally administered. In this setting legal and educational structural issues must be considered that define the powers to operate, control, and manage the schools (Alexander & Alexander, 1992).

In analyzing the American educational system and comparing it to central state systems of education in foreign countries, one is struck by the diversity of authority under which the American public schools are governed. As a federal and not a national system, the government of the United States comprises a union of states united under one central government. The particular form of American federalism creates a unique educational system, which is governed by laws of fifty states with component parts amounting to several thousand local school district operating units. Through all of this organizational multiformity, and indeed complexity, runs a legal basis on which the entire system is founded, those generally prescribed by our constitutional system.

The critical position of education in a democratic society is self-evident. Over the years the courts have come to conclude that society is best served by an educational system that teaches "through wide exposure to that robust exchange of ideas which discovers truth out of multitude of tongues. Thus because of the importance of the schools and because this robust exchange of ideas is vital to the educational process, the perpetuation of that exchange is, at all levels of the educational system, a special concern of the First Amendment" (Alexander & Alexander, 1992, p.229). No school can function appropriately as a place for the exchange of ideas unless both students and faculty enjoy an atmosphere conducive to debate and scholarly inquiry.

With this in mind, the reform of public education which includes the improvement of educational programs for those children who are least well off, should remain open to alternative views and divergent conceptions of evaluation. Critical evaluation can begin to provide an accurate analysis of the production of inequality or the reproduction of social injustice in the public schools. The ideology of critical evaluation can begin to influence a movement toward the realization of an egalitarian ideal and the elimination of inequality. I have asked educators and evaluators of education reform efforts to reconsider critically their roles in social science research, to reclaim the battleground of public school reform by focusing on the democratic purpose of public schooling, and the institutional problems in educational programs and practice that often inhibit action toward this ideal.

Conclusion

Irrespective of the many social, economic, technological, cultural, and political problems that face our American communities, the public schools exist for the purpose of educating all children. Teachers are a part of the never-ending struggle to create conditions in which learning takes place and provide the best educational opportunities in a given situation. As evaluators rendering judgement on educational programs, and giving merit or not giving merit to the educational repertoires and learning outcomes of teachers; we also become inextricably linked to the process of either perpetuating an already neutral disconnected reality of education or critically examining and observing a wide range of crucial issues, structures, and problems in contemporary education. As evaluators of education programs and teaching, we cannot ignore that we become a part of the never-ending struggle to make judgment calls about a social activity which creates the conditions or obstacles for social mobility.

The central task of the current reform movement in education is nothing less than building and transforming schools that are struggling to achieve democratic ideals (Fine, 1994). While schools can be described as potentially a site of extraordinary democracy, the processes and outcomes of schools deeply reproduce and promote the very social inequities they are said to equalize (Fine, 1983). This circumstance imposes onto the roles of educational leaders and critical evaluators a social responsibility, one that demands sincere conscience and deliberate action. Evaluators and researchers, who in the past have been content to describe dispassionately what schools are doing and how they are functioning, are actually involved in and committed to a collaborative view of knowledge creation. These characters in social change should not struggle to find a voice that sensitively captures both the insider's and outsider's view of reality. When characters, such as evaluators of educational reform, gain the conscience and purposefulness of their critical role, no relationship is left untouched or unchanged.

> In conclusion, evaluation is a powerful social force that has evolved only recently in advanced capitalistic societies, a new institution that promises to be a major influence over the long term. Its influence can be both good and bad. In either case, society before formal evaluation is not the same as society afterward. Exactly what shape the practice, institution, profession, and discipline will take in the future is impossible to predict. What is clear is that its fate will be bound to the government and the economic structure and determined in part by its own history and traditions. Some of its destiny lies within the control of the evaluators themselves; some does not (House, 1993, p.172).

Note. Paper presented at the National Evaluation Conference, Youngstown State University, Youngstown, Ohio, September, 1998

References

Alexander, K., & Alexander, M. D. (1992). *American public school law*. (3rd ed.). St. Paul, MN: West Publishing Company.

Apple, M. W. (1996). *Cultural power and education*. New York: Teachers College Press.

Apple, M., and Beyer, L. (Eds.) (1988). *The curriculum: Problems, politics, and possibilities*. Albany: State University of New York Press.

Apple, M. W. & Beane, J. A. (Eds.) (1995). *Democratic education*. Washington, D.C.: Association for Supervision and Curriculum Development.

Apple, M. W. & Carlson, D. (Eds.) (in press). *Critical educational theory in unsettling times*. Boulder, CO: Westview Press.

Arreaga-Mayer, C., & Greenwood, C.R. (1986). Environmental variables affecting the school achievement of culturally and linguistically different learners: An instructional perspective. NABE: *The Journal for the National Association for Bilingual Education, 10* (2), 113-135.

Barth, R. S. (1986). On sheep and goats and school reform. *Phi Delta Kappan, 68* (4): 293-296.

Berliner, D.C., & Biddle B. J. (1997). *The manufactured crisis: Myths, fraud, and the attack on America's public schools*. White Plains, NY: Longman Publishers USA.

Birdsall, N., & Hecht R. (1995). *Swimming against the tide: Strategies for improving health.* Working Paper No. 305, Washington, D. C.: Inter-American Development Bank, Office of the Chief Economist.

Burke, K. (1966). *Language as symbolic action: Essays on life, literature, and method.* Berkeley: University of California Press.

Coleman, J. (1975). Racial segregation in the schools: New research with policy implications. *Phi Delta Kappan, 57* October 1975, 75-78.

Coleman, J. (1981). Quality and equality in American education: Public and catholic schools. *Phi Delta Kappan, 63*, 159-164.

Chelimsky, E., & Shadish, W.R. (Eds.) (1997). *Evaluation for the 21st century: A handbook*. Thousand Oaks, CA: SAGE Publications.

Cousins, J. B, Donahue, J. J., & Bloom, G. (1996). *Understanding collaborative evaluation: Results from a survey of North American evaluators*. Unpublished paper submitted for publication, University of Ottawa.

Cronbach, L. J., & Associates. (1980). *Toward reform of program evaluation*. San Francisco, CA: Jossey-Bass.

Dewey, J. (1944). *Democracy and education*. New York: The Free Press.

Donmoyer, R. (n.d.). *Evaluation as deliberation: Theoretical and empirical explorations*. (Grant No. G 810083). Ohio State University, Columbus, National Institute of Education.

Fetterman, D. (1981). Blaming the victim: The problem of evaluation design and federal involvement, and reinforcing world views in education. *Human Organization, 40* (1), 67-77.

Fetterman, D. (Ed.). (1988). *Qualitative approaches to evaluation in education: The silent scientific revolution*. New York, NY: Praeger Publishers.

Fetterman, D. & Pitman, M. A. (1986). *Educational evaluation: Ethnography in theory, practice, and politics*. Beverly Hills, CA: SAGE Publications.

Fine, M. (1983). Perspectives on inequality: Voices from urban schools. In L. Bickman, ed. *Applied Social Psychology Annual IV*. Beverly Hills: SAGE Publications.

Fine, M. (1983b). Dropping out of high school: The ideology of school and work. *Journal of Education, 165*, 259-272.

Fine, M. (1988). De-institutionalizing educational inequity. In Council of Chief State School Officers. (Eds.). *At risk youth: Policy and Research*. New York: Harcourt Brace Jovanovich.

Fine, M. (Ed.). (1994). *Chartering urban school reform: Reflections on public high schools in the midst of change*. New York: Teachers College Press.

Fine, M., & Weis, L. (Eds.). (1993). *Beyond silenced voices: Class, race, and gender in United States schools*. Albany, NY: State University of New York Press.

Freire, P. (1970b). Cultural action and conscientization. *Harvard Educational Review, 40* (3), 452- 477.

Freire, P. (1973). *Education for critical consciousness*. New York:

Seabury.

Freire, P. (1985). *The politics of education*. South Hadley, MA: Bergin and Garvey.

Giroux, H. A. (1997). *Pedagogy and the politics of hope: Theory, culture, and schooling*. Boulder, CO: Westview Press.

Giroux, H. A. (1983). *Theory and resistance in education: A pedagogy for the opposition*. South Hadley, MA: Bergin and Garvey.

Giroux, H. A. (1983b). *Critical theory and educational practice*. Geelong: Deakin University Press.

Giroux, H. A. (1988). *Teachers as intellectuals: Toward a critical pedagogy of learning*. South Hadley, MA: Bergin and Garvey.

Giroux, H. A. & Aronowitz. S. (1991). *Postmodern education: Politics, culture, and social criticism*. Minneapolis, MN: University of Minnesota Press.

Gross, S. (1993). Early mathematics performance and achievement: Results of a study within a large suburban school system. *Journal of Negro Education, 62* (3), 269-287.

Guba, E. G. (1965). *Evaluation in field studies*. Address at evaluation conference sponsored by the Ohio State Department of Education, Columbus.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: SAGE Publications.

Habermas, J. (1971). *Knowledge and human interest*. Translation. Boston: Beacon Press.

Habermas, J. (1981). *The theory of communicative action, volume 1: Reason and the rationalization of society*. Boston: Beacon.

Habermas, J. (1987). *The theory of communicative action, volume 2: Lifeworld and system, a critique of functionalist reason*. Boston: Beacon.

Habermas, J. (1990). *Moral consciousness and communicative action*. Cambridge: MIT Press.

Hamnett, M. P., Kumar, K., Porter, D. J., & Singh, A. (1984). *Ethics, politics, and international social science research: From critique to praxis*. East-West Center: University of Hawaii Press.

House, E. R. (1976). Justice in evaluation. In G.V. Glass (Ed.), *Evaluation studies review annual* (Vol. 1, pp.75-100). Beverly Hills, CA: SAGE Publications.

House, E. R. (1980). *Evaluating with validity*. Beverly Hills, CA: SAGE Publications.

House, E. R. (1988). *Jesse Jackson and the politics of charisma: The rise and fall of the PUSH/Excel program*. Boulder, CO: Westview.

House, E. R. (1990). Methodology and justice. In K. A. Sirotnik (Ed.), *New directions for program evaluation, Vol.45*, (pp. 23-36). San Francisco: Jossey-Bass.

House, E. R. (1993). *Professional evaluation*. Newbury Park, CA: SAGE Publications.

Howard, G. S. (1985). The role of values in the science of psychology. *American Psychologist, 40*, 255-265.

Kanpol, B. (1997). *Issues and trends in critical pedagogy*. Cresskill, NJ: Hampton Press.

Kohlberg, L. (1984). *Essays on moral development: The psychology of moral development*. Vol.2. New York: Harper & Row.

Kourilsky, M. & Baker, E. (1976). An experimental comparison of interaction, advocacy, and adversary evaluation. *Center on Evaluation, Development, and Research (CEDR) Quarterly, 9*, 4-8.

Kozol, J. (1991). *Savage Inequalities*. New York: Crown Publishers.

Kretovics, J., Farber, K., & Armaline, W. (1991). Reform from the bottom up: Empowering teachers to transform schools. *Phi Delta Kappan, 73* (4), 295-299.

Kretovics, J., & Nussel, E. J. (1994). *Transforming urban education*. Needham Heights, MA: Allyn and Bacon.

Kuhn, T. (1970). *The structure of scientific revolutions*. 2nd Ed. Chicago: University of Chicago Press.

Levine, D. U. (1971). Concepts of bureaucracy in urban school reform. *Phi Delta Kappan, 52*, 329-333.

Levine, M. (1974). Scientific method and the adversary model. Some preliminary thoughts. *American Psychologist, 29*, 661-677.

Lincoln, Y. S. (Ed.). (1986). *Organizational theory and inquiry: The paradigm revolution*. Newbury Park, CA: SAGE Publications.

Lincoln, Y. S. (1991). The arts and sciences of program evaluation. *Evaluation Practice, 12* (1), 1-7.

Madaus, G. F., West, M.M., Harmon, M.C., Lomax, R G., & Viator, K.A. (1992). *The influence of testing on teaching math and science in grades 4-12*. Boston: Boston College Center for the Study of Testing, Evaluation, and Education Policy.

Maruyama, G. M., & Deno, S. (1992). *Research in educational settings*. Newbury Park, CA: SAGE Publications.

McCormick, L., Haring, N. G., & Haring, T.G. (1990). *Exceptional children and youth*. (6th ed.). New York, NY: Macmillan Publishing Company.

McLaughlin, M. W. (1975). *Evaluation and reform: The elementary and secondary education act of 1965*. Cambridge, MA: Ballinger.

MacPherson, C. B. (1987). *The rise and fall of economic justice*. Oxford, UK: Oxford University Press.

Miller, S. I., and Safer, L. A. (1993). Evidence, ethics, and social policy dilemmas. *Education Policy Analysis Archives, 1* (9) (Available online at http://epaa.asu.edu)

Murray, C. A. (1983). Stakeholders as deck chairs. In A. Bryk (Ed.), *New directions for program evaluation: Vol. 17*. Stakeholder-based evaluation (pp.58-61). San Francisco: Jossey- Bass.

Murray, C. A. (1984). *Losing ground: American social policy, 1950-1980*. New York: Basic Books.

National Council on Education Standards and Testing (NCEST). (1992). *Raising standards for American education: A report to Congress, the secretary of education, the national education goals panel, and the American people*. Washington, D.C.: National Council on Education Standards and Testing.

Nevo, D. (1986). The conceptualization of educational evaluation: An analytical review of the literature. In E. R. House (Ed.), *New directions in educational evaluation*.(pp. 15-29). London, UK: Falmer Press.

Noll, J. W. (1997). *Taking Sides: Clashing views on controversial education issues*. (9th ed). Guilford, CT: McGraw-Hill.

Oakes, J. (1986). Tracking, inequality, and the rhetoric of reform: Why schools don't change. *Journal of Education, 168* (1), 60-79.

Oakes, J. & Sirotnik, K. A. (Eds.). (1986). *Critical perspectives on the organization and improvement of schooling*. Hingham, MA: Kluwer-Nijhoff.

Oakes, J., & Guiton, G. (1995). Opportunity to learn and conceptions of educational equality. *Educational Evaluation and Policy Analysis, 17* (3), 323-336.

Ogbu, J. U., & Matute-Bianchi, M. E. (in press). Understanding socio-cultural factors: Knowledge, identity, and school adjustment. In California State Department of Education (Ed.), *Socio-cultural factors and minority student achievement*. Sacramento: Author.

Owens, T. R. (1973). Educational evaluation by adversary proceeding. In E. R. House (Ed.), *School evaluation: The politics and process*. Berkeley, CA: McCutchan.

Patton, M. Q. (1994). Developmental evaluation. *Evaluation Practice, 15*, 311-319.

Patton, M. Q. (1997). *Utilization Focused Evaluation* (3rd ed.). Thousand Oaks, CA: SAGE Publications.

Paul, S. (1991). *Accountability in public services: Exit, voice, and capture*. Washington, D.C.: The World Bank.

Prilleltensky, I. (1994). *The morals and politics of psychological discourse and the status quo*. Albany: State University of New York Press.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Belknap Press.

Rossi, P. H, & Freeman, H. E. (1993). *Evaluation: A systemic approach* (5th ed.). Newbury Park, CA: SAGE Publications.

Sabia, D. R., Wallulis, J. (Eds.). (1983). *Changing social science: Critical theory and other critical perspectives*. Albany, NY: State University of New York Press.

Schön, D. A. (1983). *The reflexive practitioner: How professionals think in action*. New York: Basic Books.

Scriven, M. (1986). Evaluation as a paradigm for educational research. In E. House (Ed.), *New directions in educational evaluation*. (pp.53-67). Philadelphia, PA: Farmer Press.

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: SAGE Publications.

Scriven, M., & Kramer, J. (January 1995). Risks, rights, and responsibilities in evaluation. *Australasian Journal of Evaluation 15*, 15-19.

Sechrest, L. (1992). Roots: Back to our first generations. *Evaluation Practice, 13* (1), 1-7.

Shadish, W. R. Jr., & Epstein, R. (1987). Patterns of program evaluation practice among members of the evaluation research society and evaluation network. *Evaluation Review, 11*, 555-590.

Shapiro, H. S., & Purpel, D. E. (1998). *Critical social issues in American education.* (2nd Ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Sirotnik, K. A. (Spring, 1990). Evaluation and social justice: Issues in public education. In K. A. Sirotnik (Ed.), *New directions for program evaluation, Vol.45*. San Francisco: Jossey-Bass.

Sirotnik, K. A. & Oakes, J. (Spring, 1990). Evaluation as Critical Inquiry: School improvement as a case in point. In K. A. Sirotnik (Ed.), *New directions for program evaluation, Vol.45*. Evaluation and social justice: Issues in public education. (pp. 37-59). San Francisco: Jossey-Bass.

Sizer, T. (1984). *Horace's compromise: The dilemma of the American high school.* Boston, MA: Houghton Mifflin.

Stake, R. E. (1986). *Quieting reform: Social science and social action in an urban youth reform*. Champaign: University of Illinois Press.

Stake, R. E., & Gjerde, C. (1974). An evaluation of T- CITY, The Twin City Institute for Talented Youth. In R. H. P. Kraft, L. M. Smith, R. A. Pohland, C. J. Brauner, & C.Gjerde (Eds.), *Four evaluation examples: Anthropologist, economic, narrative, and portrayal* (AERA Monograph Series on Curriculum Evaluation No. 7). Chicago: Rand McNally.

Tsoi Hoshmand, L. L. (1994). *Orientation to inquiry in a reflective professional psychology*. Albany: State University of New York Press.

Texas Education Agency. (1996). *Texas School Law Bulletin*. Austin, TX: West Publishing Company.

Vervoort, C. E. (1975). Onderwijs en maatschappelijke ongelijkheid als verdelingsprobleem (Education and social inequality as a distribution problem). In P. van der Kley & A. A. Wesselingh (Eds.). *Onderwijs en maatschappelijke ongelijkheid.* Boekaflevering Mens en Maatschappij, Jrj, 50. Rotterdam: UPR.

Visions of public service. (1986, Fall/Winter). *JFK School of Government Bulletin.*

Warren, R. L. (1963). *Social research consultation*. New York: Russel Sage.

Wertsch, J. V. (1998). *Mind as action*. Oxford: Oxford University Press.

Wesselingh, A. (1997). Spheres of justice: The case of education. *International Studies in Sociology of Education*, 7(2), 181-194.

Wexler, P. (1983). *Critical social psychology*. Boston, MA: Routledge & Kegan Paul.

Winfield, L. F. (1993). Investigating test content and curriculum content overlap to assess opportunity to learn. *Journal of Negro Education, 62* (3), 288-310.

Worthen, B. R., Sanders, J. R., & Fitzpatrick, J. L. (1997). *Program evaluation: Alternative approaches and practical guidelines*. (2nd ed.). White Plains, NY: Longman Publishers.

Young, R. E. (1990). *A critical theory of education: Habermas and our children's future*. New York: Teachers College Press

Note: I would like to thank Dr. Patricia Whang and Dr. Cynthia Reed for their advice and support during the writing of this article.

About the Author

Gisele A. Waters

Educational Foundations Leadership and Technology
3084 Haley Center
College of Education
Auburn University
Auburn, AL 36849

Email: waterga@mail.auburn.edu

Education:

Student studying for Ph.D. in Educational Psychology
M.Ed. Special Education/Educational Psychology, 1996. University of Houston
B.A. Economics, 1990. University of Texas at Austin
Certification: Texas; Generic Special Education (K-12),
Bilingual and ESL Education, Elementary Education

Working in Texas as a public school teacher for six years in bilingual/ESL education classrooms, and special education resource significantly influenced my critical analysis of the educational process. During that time, I held delegate positions of leadership on district and campus advisory teams, sharing the responsibility in the development and implementation of district and campus level improvement plans. The obstacles encountered to implement research based best practices carefully positioned my observations and questions about teaching and learning. Ultimately, my heart lies with teachers and children and the institutional pressures that affect them. Today my teaching at the undergraduate level reflects an instructional approach that frames my passions for issues of social justice, democracy, power, voice, and equity in schools and schooling. My methodological interests lie in the cultivation of a critical social science, a science intended to empower those involved to change as well as to understand their world.

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey@olam.ed.asu.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

Greg Camilli
Rutgers University

John Covaleskie
Northern Michigan University

Andrew Coulson
a_coulson@msn.com

Alan Davis
University of Colorado, Denver

Sherman Dorn
University of South Florida

Mark E. Fetler
California Commission on Teacher Credentialing

Richard Garlikov
hmwkhelp@scott.net

Thomas F. Green
Syracuse University

Alison I. Griffith
York University

Arlen Gullickson
Western Michigan University

Ernest R. House
University of Colorado

Aimee Howley
Ohio University

Craig B. Howley
Appalachia Educational Laboratory

William Hunter
University of Calgary

Richard M. Jaeger
University of North Carolina--Greensboro

Daniel Kallós
Umeå University

Benjamin Levin
University of Manitoba

Thomas Mauhs-Pugh
Green Mountain College

Dewayne Matthews
Western Interstate Commission for Higher Education

William McInerney
Purdue University

Mary McKeown-Moak
MGT of America (Austin, TX)

Les McLean
University of Toronto

Susan Bobbitt Nolen
University of Washington

Anne L. Pemberton
apembert@pen.k12.va.us

Hugh G. Petrie
SUNY Buffalo

Richard C. Richardson
Arizona State University

Anthony G. Rud Jr.
Purdue University

Dennis Sayers
Ann Leavenworth Center for Accelerated Learning

Jay D. Scribner
University of Texas at Austin

Michael Scriven
scriven@aol.com

Robert E. Stake
University of Illinois--UC

Robert Stonehill
U.S. Department of Education

Robert T. Stout
Arizona State University

---

archives | abstracts | editors | board | submit | comment | subscribe | search

EPAA • enter the archives • browse the abstracts • the editors • the edit board
• submit article • submit commentary • search • subscribe
volume: • 1 • 2 • 3 • 4 • 5 • 6 • 7

This article has been retrieved 633 times since December 10, 1998

Education Policy Analysis Archives

Volume 6 Number 21 **December 10, 1998** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal.
Editor: Gene V Glass Glass@ASU.EDU.
College of Education
Arizona State University, Tempe AZ
85287-2411

Copyright 1998, the EDUCATION POLICY ANALYSIS ARCHIVES. Permission is hereby granted to copy any article provided that EDUCATION POLICY ANALYSIS ARCHIVES is credited and copies are not sold.

Boundary Breaking:
An Emergent Model for Leadership Development

Charles Webber
University of Calgary

Jan Robertson
University of Waikato

Abstract

We summarize the results of a cross-cultural on-line project for graduate students in educational leadership at the University of Calgary in Canada and the University of Waikato in New Zealand. A conceptual framework for the collaborative Internet project is presented in conjunction with a summary of relevant literature and participant views of the project. Finally, the authors propose a model for on-line graduate learning in educational leadership with the following components: construction of meaning, provision of a forum for discussion, validation of personal knowledge, generative learning, formal and informal leadership, sense of community, and international perspectives.

Introduction

A key component of professional development that results in sustainable change in educators' practice is ongoing support (Joyce & Showers, 1982; McLaughlin & Yee, 1988). For classroom teachers, ongoing support can be as basic as school timetables that provide teachers with shared planning time, administrator participation in professional development activities, or participation in action research projects. For school administrators, ongoing support can take the form of mentoring programs (Goddard, 1998) or professional partnerships (Barnett, 1987; Robertson, 1995) or study groups (Bailey, 1987).

In recent years, electronic networks have emerged as a structure to link educators involved in professional growth activities in diverse settings. For example, Writers in Electronic Residence is a successful network of several years' standing that links classroom teachers and their students across Canada with well known authors via electronic mail (Note 2). School Net is a Canadian program designed to provide members of educational organizations with access to people, resources, and information that will help to promote excellence in learning (Note 3). In New Zealand, an electronic mail discussion group called Leaders-Net was established in 1997 as a way to connect principals in schools located in both rural and urban settings throughout that country (Note 4).

In the context of providing ongoing support for professional growth, this article describes a university-based initiative designed to provide individuals enrolled in similar educational leadership development courses in Canada and New Zealand with on-line opportunities to engage in substantive academic dialogue about shared interests. A conceptual framework for the university partnership is presented, followed by a description of the participants and the processes in which they were engaged. Then, the findings of a survey of participants are presented and discussed.

Conceptual Framework for the University Partnership

The cross-cultural linkage of graduate students and professors described in this paper was based on three key beliefs. First, leadership development programs at universities should contain an international component. Second, substantive learning best occurs within the context of active participation, preferably within a professional network. Finally, electronic communication has the potential to complement face-to-face interactions in university leadership development programs.

International Focus

The current practice of educational leaders occurs in a political and social milieu that transcends international borders. School administrators in the Western world share common concerns such as school reform that emphasizes concurrent and sometimes dichotomous centralization of decision making about curriculum and funding in the hands of governments, and devolution of other decisions such as staffing to school councils (Dimmock, 1993). Educators as far apart geographically as England and Tasmania are struggling to reframe accountability using "stewardship" (Radnor, Ball, & Vincent, 1998) and "neo-pluralism" (Macpherson, 1997) as related touchstones for practice. The rise in popularity of a market model of education in Canada (Fleming, 1997), the United States (Murphy, 1995; 1996), Australia (Dimmock, 1993), and New Zealand (Codd & Gordon, 1991) and even the call for "hospitality" or the creation of a safe space for children and adults to learn (Rud, 1995), share a common base in the perceived inflexibility of schools in response to students' diverse needs.

The foregoing international trends in education have led policy makers and educators alike to look beyond their own borders for information. Webber and Townsend (1998) analyzed the similarities and differences in how Canadian and Australian teachers in Alberta and New South Wales responded to government mandates to increase educator accountability through, for example, expanded student and teacher evaluation programs. They found a shared confusion among educators in both countries about the definition of educational quality, a concerted effort by governments in Alberta and New South Wales to avoid educator involvement in decision making, and a negative impact on the morale, professional growth, and career ambitions of teachers. Perhaps as a result of shared concerns about recent educational reforms, teachers around the world have looked for guidance in the work of researchers of international repute, for example, Canadians Michael Fullan (1995; 1997) and Andy Hargreaves (1997), Americans Karen Seashore Louis and Matthew Miles (1990) and Thomas Sergiovanni (1992), and British researcher David Reynolds (1997). Even politicians, in venues such as Alberta, Canada, have based their educational policy reforms on those introduced by politicians in other countries, for example, New Zealand's Sir Roger Douglas (1993).

Clearly, it is insufficient for leadership development programs to focus solely on local or national conditions which, although critical to leadership acceptance and success, may be misunderstood without a parallel exploration of international influences on policies and practices in education.

Active Participation in Professional Networks

Emergent models for professional development include a constructivist approach (Sparks, 1995; Sparks & Hirsh, 1997) in which participants individually and collectively build knowledge structures rather than simply receive information from experts. Indeed, expanding understandings of the professional growth necessary for successful school reform have led Lieberman and Grolnick (1997, p. 193) to call for professional development opportunities characterized by "a wide array of learning opportunities, engagement and commitment to inquiry, access to real problems to solve, learning that connects to ... prior experiences, [and] opportunities to work with others..." They suggest that networking can provide one such professional development opportunity. According to Lieberman and Grolnick (1997), professional networks have several important characteristics, First, networks have the potential to provide participants with venues to articulate their tacit knowledge of educational practices, thus validating what Connelly and Clandinin (1988) called personal practical knowledge and supporting the assumption of Loucks-Horsley, Harding, Arbuckle, Murray, Dubea and Williams (1987, p. 111) that "a networking approach builds the capacity of its members to identify and solve their own problems." Also, networks have a generative nature that allow learning needs to emerge prior to the development of structures; thus, the focus and structure of seminars, mentoring initiatives, and other vehicles for learning can be refined prior to their implementation. Further, professional networks provide a plethora of opportunities for individuals to emerge as formal and informal leaders with a corresponding increase in motivation to participate. Also, the very survival of professional networks depends upon collaboration among members which, in turn, may facilitate a strong sense of community that Bell (1997) stated is necessary for successful teamwork. Importantly, networks that cross organizational boundaries may foster stakeholder relationships that are mutually beneficial, egalitarian, and safe.

One description of professional networks (Loucks- Horsley et al, 1987) outlined several conditions for success. First, networks must retain a focus to maintain members' interest and involvement. Second, network members must communicate regularly or the network loses its momentum. Further, successful networks tend to be small (Significantly, this recommendation preceded the introduction of electronic networks that are large and successful; Note 5). Importantly, networks should be simple and cheap so as to retain a low cost of active participation. Finally, network members should be able to rely upon one another for information and support. This description of the conditions under which networks tend to be successful is complemented by Smith and Wigginton's (1991) description of successful networks as characterized by voluntary participation and spontaneity but "with a strong overlay of permanence and professionalism..." (p. 199) and "opportunities for teachers to develop leadership" (p. 204). The end result, according to Smith and Wigginton (1991), can be a sense among participants of being a significant part of a larger movement.

Supporting Professional Networks With Technology

Successful learning manifests itself in alterations to beliefs and practices. However, before substantive change can occur, individuals need to clarify what change will mean for them (Fullan, 1997; Hopkins, 1987). Significantly, clarification implies meaningful communication, a construct that has been central to numerous professional growth models, including clinical supervision (Acheson & Gall, 1987; Cogan, 1983; Goldhammer, Anderson, & Krajewski, 1980), peer coaching (Joyce & Showers, 1982), differentiated supervision (Glatthorn, 1984), developmental supervision (Glickman, 1981), and cognitive coaching (Costa & Garmston, 1989). Each of these models has included a description of the skills and knowledge that are required of growth facilitators.

More recently, some of the challenges posed by professional development models have emerged as significant factors in on-line professional development networks. For instance, facilitators of on-line networks need to be aware of the importance of promoting sufficient trust among participants that they will feel comfortable discussing substantive issues within the group (Farley, 1992). That is, facilitators of on-line networks need to be aware of how perceptions of them as overly controlling leaders, perhaps even censors, can prevent network members from becoming actively involved in network dialogue; this caution is of even more importance when there is a "status hierarchy" such as that which exists between university instructors and their students (Thomas, Clift & Augimoto, 1996, p. 165).

Brent (1995, p. 3) described the reactions of individuals whose words have been controlled in some way by the network facilitator or editor:

> Five hundred years of print have accustomed us to treat our words as extensions of our own identity, not to be messed with by others without our express consent nor to be inserted into others' works without acknowledgment.

Just as planners of professional development initiatives need to guard against the tendency of those involved most closely with programs to describe the results of their work in glowing terms, designers of on- line professional development programs need to beware of the hyperbole that too often characterizes reports of on- line networks (Rogers, Andres, Jacks, & Clauset, 1990). Similar cautions apply to the unrealistic expectations, held by some network participants, of immediate and substantive dialogue with colleagues from around the globe, and to the tendency to minimize the technical challenges presented by the need to integrate software, computers, modems, and server access.

On-line facilitators also must grapple with the degree to which they should structure electronic dialogue. Waugh, Levin, and Smith (1994) described how the structure of on-line networks can range from the anarchistic approach, characterized by the free flow of ideas, to highly structured models, which are easier to organize but may restrict the breadth and depth of participant dialogue. Whatever the degree of structure decided upon by facilitators and participants, Waugh, Levin, and Smith (1994) suggested that network activities have a life cycle that include observable stages: start, implementation, refinement, and closure. Further, they advised project facilitators to balance "high tech" with a "high touch" approach that acknowledges the benefits of respecting participants' needs for brevity and careful editing of electronic exchanges. Further, they described the potential for participants to establish an electronic presence that will affect their professional reputations locally and in much broader contexts, plus the importance of clear time lines for participant exchanges and the salience of promoting shared ownership of the project. To this end, Waugh, Levin, and Smith (1994) urged facilitators to strive to promote dialogue by validating the ideas of participants, posting "cheerleader" messages, and submitting dialogue summaries that lead to spin-off conversations. Further, Thomas, Clift, and Augimoto (1996) urged on-line network facilitators to respond in person and on-line to the issues and concerns that students articulated. Clearly, successful facilitation has the potential to result in what Fulk, Steinfield, Schmitz and Power (1987) described as a "social presence," or the degree to which the medium promotes personalization, warmth, sensitivity and sociability among the people involved.

Summary of the Conceptual Framework

Based on the foregoing information, the cross- cultural linkage described in this report attempted to achieve an environment characterized by the following attributes:

- Opportunities to construct personal meaning individually and collectively.
- Provision of a forum to discuss substantive leadership issues encountered in theory and practice.
- Validation of personal practical knowledge.
- A generative approach to issue identification that encouraged participants to articulate their immediate concerns and interests.
- Emergence of formal and informal leadership.
- Creation of a strong sense of community.
- Opportunities to gain international perspectives on policies and practices.
- Flexible structure
- Proactive intervention by network facilitators

These characteristics formed the basis for the cross- cultural network that is described in the next section.

Project Structure

This cross-cultural university partnership began with the authors deciding to facilitate an on-line exchange between graduate students concurrently taking educational leadership courses in Alberta, Canada, and New Zealand; the courses focused on school culture and educational review and development. This decision necessitated the development of complementary course outlines that included a common text (Hargreaves, 1997) and a common assignment that allowed students to create an electronic portfolio consisting of postings to an international electronic mail discussion group called the Change Agency Listserver (Note 6). The six New Zealand students were required by their professor to complete the electronic portfolios while the class of twenty-one Canadian graduate students were able to choose between completing an electronic portfolio and a set of article reviews; eight Canadian students elected to complete electronic portfolios, while another six students chose to read the Change Agency postings and discuss them in class and seven students decided not to participate in the on-line dialogue either by posting or reading messages.

The eight Canadian students who completed electronic portfolios consisted of one elementary school principal, three elementary school junior administrators, two secondary school junior administrators, one elementary school teacher, and one secondary school teacher. The six Canadian students who chose to read but not post messages consisted of one elementary teacher, two secondary school teachers, one secondary school junior administrator, and two instructors in postsecondary institutions. The seven nonparticipants included four elementary school teachers and three secondary school teachers. The New Zealand graduate students included two principals, one deputy principal, and three international students. The international students were a primary teacher, a College of Education president and a Ministry administrator from Zimbabwe, Solomon Islands and Indonesia.

Students in both settings who completed electronic portfolios could participate in the on-line dialogue in any of the following ways:

- Respond to e-mail messages posted by others on the Change Agency.
- Post messages to the Change Agency that did not respond to someone but started the general discussion in another direction.
- Send an e-mail message to the professor(s) and reflect on how the general discussion related to the course readings and class discussions.
- Send an e-mail message to the professor and make suggestions about how to improve the Change Agency.
- Post brief reviews of books and/or articles that related to the topic of school culture.

The messages from each student to the Change Agency or the professor were collected in an electronic file or "portfolio" and evaluated according to these criteria: breadth, depth, clarity, evidence of critical scholarship, and technical quality.

Both the Canadian and the New Zealand professors told their students that the varying levels of experience with both e-mail and teaching that students brought to the course would be recognized in the evaluation of the electronic portfolios. It was expected that students who were familiar with e-mail would contribute a larger number of postings to the Change Agency than those who had not used e-mail before the class. Further, it was anticipated that junior and senior educators would introduce different content to the electronic discussions. Finally, students were told they should expect that they would have diverse backgrounds as teachers, consultants, and administrators in a variety of educational contexts and that their diversity likely would be reflected in the electronic portfolios.

The cross-cultural graduate student exchange was limited by the fact that, due to different university schedules, the university terms overlapped by only six weeks. Therefore, the on-line exchange occurred over a six-week period, the first two weeks of which were used to get students in both settings fitted with e-mail accounts and, in some instances, to familiarize students with e-mail software and computer hardware.

An open-ended survey instrument was developed by the two researchers to determine the utility and impact of the collaborative project. It was piloted with four Canadian students who had completed electronic portfolios in previous courses with the Canadian professor, revised based on student feedback, and then administered to all students in the two courses.

The data were analyzed through a series of detailed readings to discern patterns and categories that emerged from the data rather than from a predetermined framework. The resulting categories of information are described in the following section.

Findings

This description of the results of the graduate student survey is presented as a series of categories that summarize the substance of responses to the survey items. This format was considered by the authors to be more reflective of the patterns of information provided by respondents than a summary that followed the sequence of items in the instrument would be.

Enhanced opportunities for gaining critical perspectives nationally and internationally

The opportunity to explore international influences and perspectives on policies and practices in education was one of the major themes that came through the students' responses in the survey on the collaborative study. They articulated the importance of seeing the bigger picture of issues in education internationally. As one student so aptly put it, "I think it was most powerful for me in its reminder that there's a world out there!" The listserver gave students the access to a far wider community of scholars than their usual graduate classrooms. Students from New Zealand, Zimbabwe, the Solomon Islands, Canada, and Indonesia took part in discussion and debate. One student said he would tell others about this learning experience in this way: He said he would "...strongly advise them to take part as it will help them feel connected to educators worldwide." Another said "Issues presented by our New Zealand counterparts brought forth a variety of perspectives which would otherwise have not been considered, i.e. the Zimbabwe colleague whose discussion of 'postmodernism' reminded us to look beyond our own situations to a 'global' view." One student said "I have been able to develop international perspectives through the responses of members. It gave me the opportunity to broaden my understanding of how things happen or are done in different parts of the world." The fact that there were similar issues being confronted by educators across the Commonwealth was also noted as a positive outcome of the project as it gave the students opportunities for collective construction of meaning. The outcomes from this creation of counter cultures within the learning framework were ideas and possibilities, affirmation and challenge. One student described it like this: "Good to know others are having similar experiences and what they are doing about it." This led to a feeling of global community among members of the educational leadership courses.

Developing a sense of community

The students felt they were given a unique opportunity to 'meet' with people in other parts of the world. This was something they would normally have not had the opportunity to do in their graduate studies. One thought there might be chances for study visits or sabbaticals in the future. Her concluding statement was, "In my culture there is a saying that 'those who have met, will sometimes meet again.'" This summarized the connections she had experienced with these newfound friends on the other side of the Commonwealth.

There were many interesting comments from the students about how well they felt they knew their colleagues on the other side of the world through their discussions on the listserver. The students who were in the same class on campus felt they gained new insights about their face-to-face classmates through their discussion on the listserver, but found that it was more difficult to get to know those students who were in a different classroom on the other side of the Commonwealth. Early in the project students in New Zealand asked for profiles of their Canadian counterparts as they felt they were writing to an unknown audience and initially found this difficult. They felt they needed a greater knowledge about their counterparts' interests, educational positions and professional issues. These were provided. A Canadian student, in his final evaluation, also suggested that receipt of a profile of each participant before posting began would "add interest and context to the discussion." Along the same lines, another student suggested that we "begin class with having to e-mail a classmate or you [the two professors] [a] letter of introduction."

On the other hand, others found that writing to an unknown audience made things easier. One student said, "Sometimes it is easier to say what I want to say without looking at a face." Students also found the "think time" before making a response a valuable part of taking part in the asynchronous nature of electronic discussions. One student said that the listserver discussion "allows time to hear 'their voice' and decide to agree or disagree." However, the students were also really surprised and pleased at the rapidity of responses to their contributions. One of the students commented that "whilst this was not a face-to-face communication I, however, felt as if I was talking directly to someone. Above anything else, electronic group discussions make learning fun and exciting." Another student voiced her enjoyment of the new form of communication by stating "I know how to communicate in a new way and do so daily."

Another outcome of the involvement in the collaborative study was the positive impact that it made on the complementary in-class sessions that were being held on each campus. Students were motivated and excited about responses they received on the Change Agency prior to their class sessions. One student summarized this by saying, "There was great anticipation by participants regarding how saying, "There was great anticipation by participants regarding how others would respond to their postings," and indeed disappointment when there was no response. Students felt they generally had thought in far more depth about the articles and the discussion carried out on-line, and this depth of critique carried over to their continuing class discussions. One student who saw the two types of interaction as complementary, giving him a greater understanding of his colleagues., said "This was like reading other people's papers. It allows perhaps a deeper look at your colleagues" (rather than face-to- face). Canadian students wondered whether those students in the class who didn't take part in the collaborative project felt "left a bit out of the Change Agency 'loop'."

However, some students found they enjoyed the class discussions more than the on-line discussion. One student said "I enjoyed the in-class discussions more because they involved more people than [those who] responded to a given posting." Another said, "I prefer in-class sessions because I feel a bond is easier to develop. The non-face-to-face bonds develop as well but take time." This 'bond' was also referred to by another of the students who could see good potential if time were given. She said "I believe that a bond could easily develop [among] individuals, schools and countries after the initial interaction on the Change Agency. You very quickly see someone who sums up education as you see it, who you can really relate to. I believe this is a good form of professional development." Therefore the students identified their graduate study as their professional development and saw that with the Change Agency, this could continue after the last course assignment was due.

Continuing professional development

The students involved in this collaborative study saw the electronic portfolio assessment option as more than an assignment for a course of study. The findings indicated that they could see that it could make a valuable contribution to their continuing professional development well after the assignment or course of study was over. One student summarized it this way: "I would suggest that the use of electronic portfolios is an effective way to develop one's knowledge base while gaining a very current perspective in a specific educational area. As well, this fosters critical writing and reading which will benefit the student at the conclusion of the program." This pervasive theme in the findings was also summed up by another participant who commented that "I will be able to keep current through the Change Agency after my courses are done." These students involved in graduate study talked in class about the positive effects of tertiary study on their practice and their ability to keep up with the ever developing knowledge base. They therefore valued the opportunity to establish a presence in a forum which could continue well after the graduate classes had finished. Nearly all of the students involved in the collaborative project did not unsubscribe from the listserver in the six months after the project had finished.

Continued access to research and literature, especially when studying at quite a distance from a university campus, was appreciated by some students and summed up by one student who said she found that "reading quotes from literature in others' contributions gives me a wider knowledge base than [that which] is readily available when studying at a distance."

Students used ideas from the postings with their professional colleagues outside of their university course work. One women set up a file of contributions for other school members to view. Others made statements like "[it] has also led to interesting conversations with colleagues [outside of the Change Agency]" and another said "I shared one of the postings with my colleagues and it has generated quite a discussion." Another used the contributions to generate discussion in her local principals' group. The learning community was being redefined through this process.

The fact that the "information is current, up to the minute" was noted by the students. Not only were they critiquing recent publications, but the students' contributions were written about issues of immediate concern and interest. Also, there was a number of postings on particular issues such as networking, future trends, and postmodernism, which built into a source of reference material for future use.

Influence on professional beliefs and practices

Although more difficult to ascertain from the students' responses, we believe there was an influence on students' professional beliefs and practices through the reading and posting of contributions on the listserver. One student in particular intentionally sought clarification and challenge of his own beliefs, values and practices. He said "I have tried in all my contributions and discussions to use or reflect on situations from my country in the hope that I will receive contributions or critiques from members which will help me adjust my perceptions or practices in the areas of policies and practices in education." Another student stated that this learning experience "promotes reflection and analysis of personal beliefs." Other students' responses to questions on the survey did indicate that their involvement in the collaborative study had prompted them to change their leadership practice or take action in some way. One student said "Some postings have given me metaphors that help me understand certain ways of thinking and made me reflect on my practice," and then gave specific examples of these from postings both from Canada and New Zealand. One student commented on her further reflection on the issue of student respect and the wearing of hats and said "I have really done some hard thinking and am looking at this issue with my students and parents."

An over-riding theme, through the comments the students made about their involvement on the Change Agency, was the power it had of making them reflect upon their own value positions, culture and ways of knowing. One student said "One re-examines one's own outlook through the eyes of a reader from overseas. For example, when communicating internationally one has to provide context which often simplifies our own issues." The students had to negotiate cultural boundaries in their pursuit of understanding and being understood. Another student commented about the positive challenge to critically reflect as part of this process. She said, "Critical opinions are stated in ways that are not demeaning or hurtful. I really think this helps to push the edges of our reflections of our own beliefs and practices."

Another area of influence on professional beliefs of this electronic task was that the students gained in confidence by personally using e-mail and the Internet, and several students were planning ways they would take a greater leadership role in promoting technology usage by students and teachers in their educational institutions. This was particularly true of those educators who had not used e-mail prior to this project. At the beginning of the project one student shared her "trepidation about the unknown and feeling of inadequacy" and ended the study by saying, "I'm pleased I had to do this and have found the interchange of information and exposure to the ideas of others to be very valuable and at times challenging." Significantly, one student stated that she believed that personal experience like that provided by the collaborative electronic study was imperative for an educational leader. She said, "In the future we will be using e-mail as an educational tool for our children and so we need personal experience as teachers [with] the benefits and practicalities of this process." The new skills they learned were seen as an added bonus to the benefits of being involved in the collaborative study.

Publishing Skills and Opportunities

Not only were the students challenged to reflect but the students also commented on how they had developed in being able to put forward a strong case or perspective on particular issues. They said this necessitated being able to think carefully and to make sure they had read well on the subject. One student said that responding to issues "...forced me to do additional readings on topics to expand my perspectives or to support my personal belief." Another agreed that "in making a contribution it's a real commitment of your own ideas when going public, so they have to be well founded." One student stated, "Writing for a particular audience (potentially global) in a particular format...requires a certain ability to analyze, [and] synthesize in a succinct manner." Others concurred that a short posting was much more difficult to develop than a full 3000-word assignment as they had to think more carefully about what to say when word length was limited. One student went so far as to say that compared with literature reviews and critical essays, the contributions to the listserver incorporated a process that "is superior because it broadens one's perspective so much more and one is accountable to a much wider audience." However, one student did not post to the Change Agency because "the brevity of the postings did not allow much critical discussion" and another student preferred to read the longer contributions and enjoyed the chance to read more than one response on a particular issue.

Students received international "publishing" opportunities which are seldom afforded graduate students. One student was "spotted" by the editor of an international journal who asked her to further develop her contribution and submit it to the journal. The student described this publishing experience by stating that "the opportunity has afforded me 'courage of voice' [and] stretched me into realizing the potential of shared ideas.

Opportunity for innovation and challenge

The students commented in their final course evaluations about the uniqueness of the electronic portfolio assignment as part of their graduate course work. They not only valued the opportunity to be given a variety of assessment options within the course, but also felt that they were taking part in something which was an exciting innovation. Further, they used words such as "exciting," "progressive," and "valuable" to describe the cross-cultural project. One student described the joint initiative as a "very progressive and valuable collaborative effort," and went on to say that, "the major value of this is that it is current - [happening] right now." Another student supported this by stating that she felt that is was "a real activity, [with] real people on the other end." The students felt that it was a useful activity, that "your words actually count" and that others were interested in their viewpoints. They were writing with a purpose and receiving constant feedback. It was too real for one who was "concerned about the consequences of my words on, for instance, central office!"

Some students talked about "possibilities," that is, chances to explore leadership issues collaboratively and motivation to use e-mail with the children in their classes. A sense of global community was aroused and students raised the possibility of this type of study being taken one step further and actually meeting with the students they had discussed and debated key issues with in a collaborative study tour exchange. One student said that the experience "makes me think about possibilities such as school contact with other countries, doing courses by e-mail."

Computer skill level and confidence

Computer skill level and confidence influenced student involvement in the Change Agency. The New Zealand students were not given an option for this particular assignment and their responses echoed their fears at the beginning. Many of these students said they would not have chosen to become involved but, in hindsight, were pleased they had no other option. One student said, "There was trepidation about the unknown, a feeling of inadequacy because of my lack of knowledge and skills, and bewilderment about the jargon, but [I'm] also pleased to be forced into it and looking forward to the personal growth and finding out what others seem to be so enthused and excited about." This student later said, "I'm pleased I've had to do this and have found the interchange of information and exposure to the ideas of others to be very valuable and at times challenging." Another student said "I was afraid of computers and I was ignorant of the wonders computers can do to help people to do things, especially in education. For a week or so I tried to avoid the rooms where the computers are...It was very scary indeed on the outset..." and later said, "It has been absolutely excellent and educational. There was a whole lot of things covering a wide range of topics and issues I would not have had access to if I had not joined the discussion group." Another student said that the learning experience was "...wonderful! I really wondered at first whether I should even venture to participate. Once I got through the initial technological 'glitches' and intimidation in writing my views and opinions for such a large and unknown audience I was fine. I felt proud of my accomplishment."

In fact, some of the students were able to link their experiences of the personal change process they underwent to their participation in this study to the theory. One student said, "Everard and Morris (1985, p.170) state change usually leads to temporary incompetence and that it is uncomfortable! How true!" Fullan's (1993) work on change also featured in their final reflections about the process they had gone through to take part in this study. One student said, "I do now challenge Fullan's writings - that people can't be forced to change (1993, p.22). This learning was forced in a way - if we had not changed and become e-mail users we would not have completed this section of the course." Another said "I now see why Fullan (1993, p.27) said 'Problems are our friends.' All those hassles at the start were worth it." Another student had advice for the instructors. She said, "My only suggestion is that you (both) strongly encourage students not to back out at the beginning of the course if it looks too scary."

The Canadian students were given an option and the majority of the students who chose not to be involved in the project said that either skill level or computer difficulties had influenced their choice of assignment formats. Some students had difficulty connecting to the listserver from their homes, while others had malfunctioning computers at the time. They mentioned words like "intimidating," "lack of time to learn," and "frustrating" in their justification for choosing not to be involved.

However, there also were Canadian students who had never used e-mail who took the electronic assignment option. Their comments were similar to their New Zealand counterparts. One student said "I would say it was VERY stressful and intimidating but a learning experience that I would encourage others to participate in because I learned a great deal." Another Canadian student who did take part suggested that the instructor "should make the 4th or 5th class an entire lab and mandate one small posting from each student." Another felt that they should "devote some actual class time to mock postings."

It is important to note that, despite their frustrations with learning new computer skills and the time that took, students from both countries were disappointed the project could not have continued longer and that the New Zealand students finished their contributions just when the Canadian students felt they were getting underway.

Guidelines and structure

The students in both countries sought more structure than was

The students in both countries sought more structure than was originally planned for. Students asked for suggestions and "starters" for ideas during the course and also worked on group responses in class sessions. This was more at the beginning of the study when they were unsure of what was expected. They gained in confidence after reading and contributing to the listserver. None of the students made any comment about the mandatory nature of some of the work, particularly in the New Zealand course where no options were given. Indeed, many students felt that all students taking part in the graduate course should be mandated to make at least one posting as part of an in-class session, and many commented that they felt that guidelines and more structure would have enhanced the project. One student stated that the study needed "more direction and feedback" because to him "at times it felt like 'hit and run'." Another student wished that there had been "private" messages of affirmation as there was no evaluative feedback until the end of the study and this did not help to allay fears during the initial stages of contributing. Finally, one student felt that "there needed to be more emphasis on collaboration in a conversational kind of style" and in a similar vein another student felt the contributions were a "bit dry" at times.

The paradox of the students wanting more structure and our belief of the necessity of a more fluid context for learning raised an issue that needed to be addressed. If we believed in all of the components of the learning framework, highlighted by the findings, overstructuring was the antithesis of what we were striving for. Their discomfort was indicative of how dependent some of their previous learning contexts have made them. We consciously worked to resist their attempts to have too much reliance on us. We knew the initial discomfort was essential in the development of intellectual independence and to enable the formal and informal leadership to emerge from the student group.

An Emergent Model

The results of this exploratory cross-cultural electronic partnership support the development of a structured framework for university-based educational leadership programs. Although the Internet has emerged as a free- flowing, often chaotic environment that fosters--in its positive manifestations--unrestrained creativity, it is obvious from the partnership described here that the successful use of the Internet as a teaching tool depends upon a clear understanding of the resultant changes to roles and expectations for participants, that also may have implications for more traditional approaches to leadership development. Table 1 (see next page) portrays an emergent technology-enhanced model for university- based leadership development programs. Many aspects of the model are consistent with widely used leadership development models. However, the nature of on-line instruction changes the model components in many ways, including the ease with which international delivery can occur. The model is described as "boundary breaking" because of its capacity to move learning beyond the boundaries normally imposed by cultures, roles, institutions, economics, and national borders.

Table 1
Boundary Breaking: An Emergent Model for Leadership Development

Attribute	Student Role	Instructor Role	Implications
Construction of meaning	● Rigorous reflection ● Active 'listening' ● Juxtaposition of self & others	● Examination of instructional practice ● Reduced role as information provider	● Co-learning ● Reduced hierarchy
Provision of a forum for discussion	● Challenging debate ● Public expression ● Self-evaluation ● Risk taking ● Cross-role dialogue	● 'Public teaching' ● Asynchronous communication ● Redefinition of 'courses' ● Shared evaluation	● Potential discomfort ● Technological infrastructure ● Computer skill development ● Seamless integration of technology
Validation of personal knowledge	● Exploration of practical experience ● Analysis of personal beliefs ● Articulation of assumptions	● Acceptance of practice-based knowledge ● Contextualized theory ● Critical analysis of relevant theory & research	● Confluence of theory & practice ● Reduced status differential
Generative approach to learning	● Active involvement ● Examination of personal practice ● New metaphors for practice	● Trust in process ● Reduced intervention ● Less control ● Diverse student needs for information	● Flexible course structure ● Varied evaluations ● Issue relevancy ● Contextualized participation
Formal & informal leadership	● Enhanced locus of control ● Embraced stress	● Shared leadership ● Modeled leadership ● Clarification of leadership practices	● Expanded participant profile ● Shared responsibility for learning
Sense of community	● Links to colleagues outside classes ● Consideration of 'others' ● Cross-role dialogue	● Attended to affective behaviors ● Encouragement ● Attention to safety ● Pastoral care	● Reduced teacher isolation ● Global community ● Enhanced local community
Growth of a counterculture	● Seeking cognitive dissonance ● Scrutiny of the heretofore accepted	● Imaging of alternatives ● Creating opportunities to question and imagine	● Pushing the edges of beliefs & practices ● Possibilizing
International perspectives	● Cross-cultural analysis ● Reconsideration of personal contexts	● Collaboration with compatible instructors ● Provision of materials ● Integration with local & national communities	● 'Big picture' focus ● Alternative perspectives

Construction of Meaning

The on-line leadership development model is intended to complement and not replace other activities such as face-to-face classes and seminars, principal internships, and independent scholarly research. As such, expectations for student and instructor participants in on-line learning should be consistent with normal standards for academic rigor. However, the on-line model is intended to create a context in which participants' reflections and understandings are subjected to intense analysis from several perspectives: self, local colleagues and instructors, peers in international settings, and individuals representing, for example, parents and policy makers. This juxtaposition of self and 'others' is designed to clarify personal understandings, promote active 'listening,' and create cognitive dissonance that motivates participants so that individual and collective meanings may be constructed. These alternative perspectives form part of the reflective observation and abstract conceptualisation of Kolb's (1984) learning cycle.

If substantive meaning-making is to occur, however, on-line instructors must alter some of their instructional practices. For example, courses must be reconstituted to permit active involvement by a wide range of individuals who are registered formally as students and others who participate informally as participants from the broader community. Thus, the saliency of the role of instructor-as-information-provider, of necessity, is reduced because 'others' also provide participants with access to theoretical and practical information. However, there is a corresponding increase in the importance of instructor-as- instructional-designer, able to formulate a learning environment that promotes co-learning and restricts traditional participant hierarchies.

Provision of an On-Line Discussion Forum

Expressing one's emergent understandings in an on- line forum is decidedly public, more so than what registrants in university leadership development programs usually expect to experience. Instructors should anticipate at least an initial reluctance among graduate students to post messages to an Internet forum. Nevertheless, the nature of on-line communication, and the resultant care that participants take with their public statements, enhances rather than reduces academic rigor. That is, messages tend to be subjected to extremely thorough analyses by authors prior to posting.

Instructors using an on-line delivery format for courses or modules also should be aware of the public scrutiny that awaits their own work. In the context of the Change Agency, participants include professors from several universities, policy makers, department of education personnel, teachers, and both school-based and central office administrators. Public and private assessment of instructors' work is immediate and widespread. Consequently, instructors must structure the on-line discussion forum carefully.

Other instructional considerations include a willingness to alter the definition of teaching to include asynchronous communication, which is of particular relevance to international participants operating in very different time zones. The resulting 'teaching' that can and does occur at all hours of the day and night requires instructors to develop patterns of work that allow them to fulfil their other research and service obligations. As well, graduate courses that had been taught previously as twelve three-hour meetings over a four-month period may need to be reconceptualized as an integrated package of face-to-face and virtual 'classes.' Consequently, student and instructor understandings of courses will be challenged and some discomfort may result, particularly among those expecting a 'typical' university course format. Even student evaluation will be altered because of the need for instructors to incorporate into assessment procedures the feedback that students get from other on-line participants.

It is worth noting here that the design of an on- line discussion forum, whether that be the construction of a listserver or the use of a news group, should include opportunities for graduate students to strengthen their computer skills. As well, it is critical that the integration of technology into instruction be as seamless and user-friendly as possible, regardless of the technological infrastructure that is utilized.

Validation of Personal Knowledge

Participation by non-students in on-line instructional settings may promote the inclusion of practice-based knowledge in conversations. Rather than weakening the academic rigor associated with graduate study, practical knowledge can serve as the basis for examining professional beliefs and articulating previously taken-for-granted assumptions. In fact, the wider the participant audience, the greater the likelihood that individuals will experience public challenges to unstated beliefs and assumptions, something that most graduate programs strive to include.

It was obvious in the project described in this report that theoretical and empirical perspectives can be integrated into dialogue as significant issues emerge from practice-oriented conversations. However, this may only be possible when university instructors are able to recognize the value of practical knowledge as a vehicle for contextualizing the academic focus of leadership development programs. That is, the on-line instructional framework proposed here may provide enhanced opportunities for theory and practice to converge in ways that are meaningful for participants who are willing to accept and, in fact, to seek a reduction to the status differential that too often is associated with theory and practice. Theory takes on a wider meaning through this validation of personal theories.

A Generative Approach to Learning

The proposed model for on-line graduate learning is generative in nature. That is, it is based on the belief that active engagement in personally meaningful activities is essential to significant learning.

Feedback from participants in the Canada-New Zealand collaborative project included mention of how the on-line dialogue made course content more meaningful because the topics of conversation emerged from individuals' professional and cultural contexts. However, participants highlighted the fact that feedback from colleagues in very different settings elicited examinations of personal practice and the construction of new metaphors that were useful frameworks for considering their professional environments.

From an instructional perspective, generativity requires a relinquishing of some control and a sufficient trust in the ability of other participants in the on-line learning community to pose and respond to learning challenges relevant to the course. Further, instructor interventions in the conversation should be relatively minimal and particularly strategic when compared to many face-to-face interactions. Conversely, there is an increase in the need for instructors to respond to a much wider range of student information needs, the result of greater individualization in course expectations.

The potential results of a generative approach to learning include a course structure that is sufficiently flexible to allow for varied evaluation formats, such as electronic portfolios and collaborative writing by students in different universities and countries. Similarly, issue relevancy may be increased and students may find that their course participation is contextualized in terms of their individual settings and in the international educational community.

Formal and Informal Leadership

Student participants in the proposed model have the potential to exercise extensive control of their learning. They are able to choose, more than in traditional courses, when they will participate in on- line dialogue and to address topics of greatest relevance to them, supported by a broad range of university- and field-based colleagues. Further, different students will emerge as dialogue leaders and information sources as different topics within the parameters of the course arise in the conversations. It is noteworthy that, despite the greater stress of learning in a significantly public setting, the strong control that students have of their learning permits them to reframe the stress so that it becomes an 'embraced stress' that is supportive and motivating.

Opportunities for shared leadership facilitate the modeling by instructors of the very leadership practices being studied by graduate students. That is, instructors can model educational leadership characterized as facilitative, collaborative, adaptive, informed, proactive, and constructive--the features of the transformational leadership so necessary in a rapidly changing, postmodern educational context. Importantly, the modeling of effective leadership practices in the context of graduate learning is enhanced if instructors describe the ways their instructional practices reflect current knowledge. That is, instructors should explain to graduate students how course organization and delivery formats were not accidental but, in fact, the result of a conscious attempt to model the manifestations of effective leadership. This clarification can emphasize to students the benefits of distinguishing between what Purkey and Novak (1984) described as intentional success versus accidental success.

The benefits of promoting formal and informal leadership within on-line graduate learning include an expanded participant profile. Several graduate students in the project described in this report found that their postings to the Change Agency resulted in requests to submit manuscripts to academic publications, to participate in policy committees within their local educational community, and invitations to apply for administrative positions. Other students reported going to meetings and seminars to discover that their postings were the basis for constructive dialogue outside of the graduate courses. These were unanticipated benefits to project participants that resulted from the public nature of their reflections, critiques of literature, and analyses of public policies.

Sense of Community

A key feature of the university-based, on-line model for graduate learning being presented is the strong sense of community that can result from its successful implementation. Project participants certainly communicated with colleagues in another country, but equally important was the fact that the electronic dialogue reflected a consideration of 'others' that increased as the project evolved. Participants found themselves responding to one another in ways that precluded the posting of treatises that were of interest only to the people writing them. In other words, participants' postings reflected considerable consideration of the needs and beliefs of the authors of other postings, including individuals in parental and policy-making roles. This became a redefining of the learning community through the dissonance created by different roles and cultures within the educational context. From this perspective, the graduate learning was strengthened by cross-cultural and cross-role dialogue that elicited a host of rich learning opportunities.

It is incumbent upon on-line instructors to attend to affective behaviors that influence a developing sense of community. That is, instructors need to send electronic mail messages to all participants in order to make gentle suggestions about posting practices that invite broader participation rather than stifle it. Alternatively, instructors should be willing to embrace the equally necessary 'invisible' work encompassed by private messages to individual participants to validate and encourage continued participation. Other behind-the- scenes work includes efforts to familiarize graduate students with computer software and hardware, and to provide them with information, authors, titles, and ideas that support students' information needs and interests.

Successful participant attention to community building can reduce teacher and administrator isolation, plus promote membership in an international community, a phenomenon that is a particularly rare experience for school-based educators. Furthermore, participants' sense of membership in a local community can be strengthened because of on-site conversations that are enriched by participation in the international on- line community.

Growth of a Counterculture

A strong counterculture emerged during the six-week cross-cultural dialogue. That is, participants quickly understood that it was safe to disagree with one another on-line and in face-to-face classes, more than might normally be expected. Students and instructors found that their taken-for-granted assumptions were challenged by participants in both countries. For example, most individuals in the two groups seemed to find relevance in discussion topics that ranged from the possibly trite-- students' wearing hats or ball caps in schools, to the seemingly narrow--the possible relationship between the length of postings and the depth of analyses, to the very broad--postmodern influences on learning. Nevertheless, virtually none of the topics featured in the on-line dialogue escaped the scepticism of some participants. Thus, cognitive dissonance was encountered by participants who previously had not considered the possibility that hat-wearing students did not necessarily lack respect for teachers, that on-line dialogue may be limited in some important ways, or that postmodern debates may have little or no relevance for colleagues in developing countries.

The emergence of a counterculture meant that the instructors had to nurture opportunities for themselves and their students to discuss how cognitive dissonance may be a prerequisite for meaningful academic discourse. The instructors learned to highlight how intended and accidental cognitive dissonance provided opportunities to embrace intellectual discomfort, rather than avoid it. Pushing the edges of beliefs and practices in this way created new possibilities for learning. We called this "possibilizing". In fact, one of the more powerful components of the learning model could be its capacity to facilitate participant understandings of the need for leaders to seek out and nurture those members of the school community who are most likely to voice uncertainty or discomfort about policies and practices. Further, the cross-cultural dialogue experienced in the present project suggests that educational leaders may benefit from developing their abilities to move comfortably and often from the perspectives of formal leaders into the worldview of a viable school counterculture.

International Perspectives

One of the most potentially beneficial components of the on-line delivery model is its international dimension. Participants in the New Zealand-Canadian project developed a deeper understanding of how educational systems in the two countries were undergoing very similar changes, often mandated by governments with equally conservative economic and political philosophies. Learning was enriched further by the views of international students, particularly those from developing countries in attendance at the New Zealand partner university, who countered the tendency of their Western colleagues to fail to consider the contexts of educational leadership in developing countries. Thus, the focus of the program promoted participants' understandings of international influences on local educational conditions.

Instructors working within the framework of the on-line model must assume responsibility for promoting the international connections necessary for successful collaboration. Responsibilities include identifying and contacting colleagues with a shared interest in collaborative on-line instruction, collectively ascertaining the compatibility of instructors and courses, and deciding upon the materials and procedures that will form the basis for the cross-cultural dialogue. Equally important is the responsibility for linking the international component of instruction to the local and national educational communities. Instructors must not neglect their responsibility to help future and present educational leaders apply their 'big picture' understandings appropriately within local communities, because a key determinant of the success of educational leaders is their ability to understand the culture of their immediate educational surroundings.

Conclusion

The proposed on-line model for graduate learning is based on understandings that emerged from a review of relevant literature and an exploratory joint project conducted in the contexts of Canada and New Zealand. However, the model should be understood to be tentative and in need of further development for several reasons. First, New Zealand and Canada share a common history in many respects: date of settlement by Europeans, cultures that evolved from British colonization in the last century, and governments that are based on the British parliamentary system. Even with strong cultural similarities, students in the two countries varied somewhat in, for example, their desired levels of formality in postings to the Change Agency listserver. Therefore, the model's applicability in countries that have greater differences remains uncertain. In addition, the model depends upon a reasonably sophisticated computerized infrastructure, something that participants in leadership preparation programs in developing countries may not be able to access easily. Even with access to advanced computers, the model depends upon participant familiarity with computer-based communications or, at least, the willingness to learn computer skills within a short time period. Moreover, differences in university timetables, exacerbated by time zone differences, restrict the degree to which collaboration can occur, particularly between universities in the northern and southern hemispheres.

Nonetheless, the proposed model has the potential to facilitate leadership development that incorporates local, national, and international interactions among educational stakeholders. This is a significant development in an era of rapid educational change influenced by factors with a global impact, particularly among Western nations. The model is based upon the concept of breaking boundaries. Cultural, political and economic boundaries were traversed. Community, institutional and role boundaries were challenged. The boundaries between theories-in-action and espoused theories were brought closer together. Technology often imposes its own boundaries, and these were overcome and fully utilised in the learning process.

A powerful description of the vision for the model came from an international student attending the New Zealand partner university: "In my view, cross-cultural and local knowledge can be enhanced by network linkages. Our electronic discussion group is part of this global/local network group...I look forward to the day when parents, teachers, and pupils in, say, Canada, New Zealand, Zimbabwe, Indonesia...would be able to share ideas making use of their experiences through the Internet." In an era when such a vision can be held by an educational leader from a developing country, perhaps the most relevant question for those of us responsible for leadership development is not "Should we adopt an on- line model for graduate learning?" but "When?"

References

Acheson, K., & Gall, M.D. (1987). *Techniques in the clinical supervision of teachers.* New York: Longman.

Bailey, A.J. (1987). *Support for school management*. London: Croom Helm.

Barnett, B.G. (1985). Peer-assisted leadership: Using research to improve practice. *The Urban Review*, 17, 47-64.

Bell, L. (1997). Staff teams and their management. In M. Crawford, L. Kydd, & C. Riches (Eds.), *Leadership and Teams in Educational Management*, (pp. 119-129). Buckingham, UK: Open University Press.

Brent, D. (1996). Epublishing and hypertext publishing. *EJournal, 6(3)* . Available on-line: http://www.hanover.edu/philos/ejournal/archive/v6n3/brent/edintro.html

Codd, J., & Gordon, L. (1991). School charters: The contractualist state and education policy. *New Zealand Journal of Education Studies, 26(1),* 21-34.

Cogan, M. (1973). *Clinical supervision*. Boston: Houghton Mifflin.

Connelly, F.M., & Clandinin, D.J. (1988). *Teachers as curriculum planners : Narratives of experience.* Toronto: OISE Press.

Costa, A., & Garmston, R. (1989). *The art of cognitive coaching: Supervision for intelligent teaching.* Sacramento, California: The Institute for Intelligent Behavior.

Dimmock, C. (1993). School-based management and linkage with the curriculum. In C. Dimmock (Ed.), *School-based management and* *school effectiveness* (pp. 1-21). London: Routledge.

Douglas, R. (1993). *Unfinished business*. New York: Random House.

Everard, K.B., & Morris, G. (1985). *Effective school management*. London: Paul Chapman.

Farley, L. (1992) Making sense of change: Strategies for educational technologists. *The Computing Teacher*, 19(7), 8-10.

Fleming, T.G. (1997). Provincial initiatives to restructure Canadian school governance in the 1990s. *Canadian Journal of Educational Administration and Policy*, 11. Available on-line: http://www.umanitoba.ca/publications/cjeap/abbre1.htm

Fulk, J., Steinfield, C., Schmitz, J., & Power, F. (1987). A social information processing model of media use in organizations. *Communication Research, 14*, 529-552.

Fullan, M. (1997). Emotion and hope: Constructive concepts for complex times. In A. Hargreaves (Ed.), *Rethinking Educational Change with Heart and Mind*, (pp. 216-233). Alexandria, VA: Association for Supervision and Curriculum Development.

Fullan, M. (1997). Planning, doing, and coping with change. In. A. Harris, N. Bennett, & M. Preedy (Eds.), *Organizational Effectiveness and Improvement in Education*, (pp. 205-215). Buckingham, UK: Open University Press

Fullan, M. (1995). Contexts: Reflections and Implications. In M.J. O'Hair & S.J. Odell (Eds.), *Educating Teachers for Leadership and Change* (pp. 66-70). Thousand Oaks, California: Corwin Press.

Fullan, M. (1993). *Change forces: Probing the depths of educational reform*. London: Falmer Press.

Glatthorn, A. (1984). *Differentiated supervision*. Alexandria, VA: Association for Supervision and Curriculum Development.

Glickman, C. (1981). *Developmental supervision: Alternative practices for helping teachers improve instruction*. Alexandria, VA: Association for Supervision and Curriculum Development.

Goddard, T. (1998). Croaks from the lily pad: Towards the provision of a peer mentoring program for principals. *International Electronic Journal for Leadership in Learning*, 2(1). Available on-line: http://www.acs.ucalgary.ca/~iejll/

Goldhammer, R., Anderson, R., & Krajewski, R. (1980). *Clinical supervision* (2nd ed.). New York: Holt, Rinehart and Winston.

Grace, G. (1995). *School leadership beyond educational management.* London: Falmer Press.

Hargreaves, A. (1997). *Rethinking educational change with heart and mind.* Alexandria, VA: Association for Supervision and Curriculum Development.

Hargreaves, A. (1997). Rethinking educational change: Going deeper and wider in the quest for success. In A. Hargreaves (Ed.), *Rethinking Educational Change with Heart and Mind,* (pp. 1-26). Alexandria, VA: Association for Supervision and Curriculum Development.

Hopkins, D. (1987). Teacher research as a basis for staff development. In M.F. Wideen & I. Andrews (Eds.), *Staff Development for School Improvement,* (pp. 111-128). Philadelphia, PA: Falmer Press.

Joyce, B., & Showers, B. (1982). The coaching of teaching. *Educational Leadership,* 40(1), 4-11.

Kolb, D.A. (1984). *Experiential learning: Experience as the source of learning and development.* New Jersey: Prentice Hall.

Leiberman, A., & Grolnick, M. (11997). Networks, reform, and the professional development of teachers. In A. Hargreaves (Ed.), *Rethinking Educational Change with Heart and Mind,* (pp. 192-215). Alexandria, VA: Association for Supervision and Curriculum Development.

Loucks-Horsley, S., Harding, C.K., Arbuckle, M.A., Murray, L.B., Dubea, C., & Williams, M.K. (1987). *Continuing to learn: A guidebook for teacher development.* Oxford, Ohio: National Staff Development Council.

Louis, K.S., & Miles, M.B. (1990). *Improving the urban high school.* New York: Teachers College Press.

Macpherson, R.J.S. (1997). Learning accountability in Tasmania: The move from command to neo-pluralist politics. *International Electronic Journal for Leadership in Learning,* 1(5). Available on-line: http://www.acs.ucalgary.ca/~iejll/

McLaughlin, M.W., & Yee, S.M. (1988). School as a place to have a career. In A. Lieberman (Ed.). *Building a professional culture in schools.* (pp. 23-44). New York: Teachers College Press.

Murphy, J. (1996). Why privatization signals a sea change in schooling. *Educational Leadership,* 54(2), 60-62.

Murphy, J. (1995). Changing role of the teacher. In M.J. O'Hair & S.J. Odell (Eds.). *Educating Teachers for Leadership and Change.* (pp. 311-326). Thousand Oaks, California: Corwin Press.

Purkey, W.W., & Novak, J.M. (1984). *Inviting school success.* Belmont, California: Wadsworth Publishing.

Radnor, H.A., Ball, S.J., & Vincent, C. (1998). Local educational governance, accountability, and democracy in the United Kingdom. *Educational Policy, 12(1-2),* 124-137.

Reynolds, D. (1997). Linking school effectiveness knowledge and school improvement practice. In A. Harris, N. Bennette, & M. Preedy (Eds.). *Organizational Effectiveness and Improvement in Education.* (pp. 251 - 260). Buckingham, UK: Open University Press.

Robertson, J. M. (1995). *Principals' Partnerships: An action research study on the professional development of New Zealand school leaders.* Unpublished Doctoral Thesis. Hamilton: University of Waikato.

Rogers, A., Andres, Y., Jacks, M., & Clauset, T. (1990). Telecommunications in the classroom: Keys to successful telecomputing. *The Computing Teacher, 17(8),* 25-28. Available on-line: http://lrs.ed.uiuc.edu/Guidelines/RAJC.html

Rud, A.G. (1995). Learning in comfort: Developing an ethos of hospitality in education. In J.W. Garrison and A.G. Rud Jr. (Eds.). *The Educational Conversation: Closing the Gap* (pp. 119-128). Albany, NY: State University of New York Press.

Sergiovanni, T. J. (1992). Why we should seek substitutes for leadership. *Educational Leadership, 49(5),* 41-45.

Smith, H., & Wigginton, E. (1991). Foxfire teacher networks. In A. Lieberman & L. Miller (Eds.). *Staff development for the 90s: New demands, new realities, new perspectives.* (pp. 193-220). New York: Teachers College Press.

Sparks, D. (1994). A paradigm shift in staff development. *Journal of Staff Development, 15(4),* 26-29.

Sparks, D., & Hirsh, S. (1997). *A new vision for staff development.* Alexandria, VA: Association for Supervision and Curriculum Development.

Thomas, L., Clift, R.T., & Augimoto, T. (1996). Telecommunication, student learning, and methods instruction: An exploratory investigation. *Journal of Teacher Education, 46(3),* 165-175.

Waugh, M., Levin, J.A., & Smith, K. (1994). Organizing electronic network-based instructional interactions: Successful strategies and tactics. *The Computing Teacher, 21(5),* 21-22 & 21(6), 48-50.

Webber, C.F., & Townsend, D. (1998). The comparative politics of accountability of New South Wales and Alberta. *Educational Policy*, 12(1-2), 177-190.

Notes

1. An earlier version of this paper was presented at the annual meeting of the American Educational Research Association in San Diego, California, April 13 - 17, 1998.
2. Writers in Electronic Residence: http://www.wier.yorku.ca/WIERHome.html
3. School Net: http://www.schoolnet.ca/info/)
4. Leaders-Net: http://www.soe.waikato.ac.nz/elc/network.html
5. American Educational Research Association Listservers: http://aera.net/resource/
6. The Change Agency Network: http://www.acs.ucalgary.ca/~cll/CAN/frameset.htm

About the Authors

Charles F. Webber
Faculty of Education
University of Calgary
2500 University Drive NW
Calgary, Alberta, Canada, T2N 1N4

Telephone: (403) 220-5694
Fax: (403) 282-8479

E-Mail: cwebber@ucalgary.ca

Web page: http://external.educ.ucalgary.ca/academic/cwebber.html

Charles F. Webber is an Associate Professor in the Faculty of Education at the University of Calgary. His teaching and research focus on school culture, the role of the principal, and international trends in educational leadership. He facilitates the Change Agency Network http://www.acs.ucalgary.ca/~cll/CAN/frameset.htm and edits the International Electronic Journal for Leadership in Learning http://www.acs.ucalgary.ca/~iejll/

Jan M. Robertson
School of Education
University of Waikato
Private Bag 3105
Hamilton
New Zealand

Telephone: 64 7 838 4500
Fax: 64 7 838 4555

E-Mail: jan@waikato.ac.nz

Jan M Robertson is the Director of the Educational Leadership Centre http://www.soe.waikato.ac.nz/elc and a Senior Lecturer in the Professional Studies Department at the University of Waikato, Hamilton, New Zealand. Her teaching, research and development focuses on action research methodology, appraisal and leadership development of teachers and principals, and the role of the principal in site-based management nationally and internationally. She initiated a listserver discussion group primarily for networking New Zealand school leaders http://www.soe.waikato.ac.nz/elc/electnetwork.html based on the Change Agency model.

Copyright 1998 by the *Education Policy Analysis Archives*

The World Wide Web address for the *Education Policy Analysis Archives* is http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692). The Book Review Editor is Walter E. Shepherd: shepherd@asu.edu . The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

EPAA Editorial Board

Michael W. Apple
University of Wisconsin

Greg Camilli
Rutgers University

John Covaleskie
Northern Michigan University

Andrew Coulson
a_coulson@msn.com

Alan Davis
University of Colorado, Denver

Sherman Dorn
University of South Florida

Mark E. Fetler
California Commission on Teacher Credentialing

Richard Garlikov
hmwkhelp@scott.net

Thomas F. Green
Syracuse University

Alison I. Griffith
York University

Arlen Gullickson
Western Michigan University

Ernest R. House
University of Colorado

Aimee Howley
Ohio University

Craig B. Howley
Appalachia Educational Laboratory

William Hunter
University of Calgary

Richard M. Jaeger
University of North Carolina--Greensboro

Daniel Kallós
Umeå University

Benjamin Levin
University of Manitoba

Thomas Mauhs-Pugh
Green Mountain College

Dewayne Matthews
Western Interstate Commission for Higher Education

William McInerney
Purdue University

Mary McKeown-Moak
MGT of America (Austin, TX)

Les McLean
University of Toronto

Susan Bobbitt Nolen
University of Washington

Anne L. Pemberton
apembert@pen.k12.va.us

Hugh G. Petrie
SUNY Buffalo

Richard C. Richardson
Arizona State University

Anthony G. Rud Jr.
Purdue University

Dennis Sayers
Ann Leavenworth Center for Accelerated Learning

Jay D. Scribner
University of Texas at Austin

Michael Scriven
scriven@aol.com

Robert E. Stake
University of Illinois--UC

Robert Stonehill
U.S. Department of Education

Robert T. Stout
Arizona State University

David D. Williams
Brigham Young University

archives | abstracts | editors | board | submit | comment | subscribe | search